REMOVAL OF NOISE FROM A VOICE SIGNAL BY SYNTHESIS

Neil Joseph Miller

Utah University

Prepared for:

Rome Air Development Center
Advanced Research Projects Agency

May 1973

128065 Removal of noise from a voice signal by synthesis

NEIL JOSEPH MILLER

UNIVERSITY OF UTAH

MAY 1975
UTEC-CSc-74-013
COMPUTER SCIENCE, UNIVERSITY OF UTAH
SALT LAKE CITY, UTAH 84112

REMOVAL OF NOISE FROM A VOICE SIGNAL BY SYNTHESIS

by

Neil Joseph Miller

May 1973

UTEC-CSc-74-013

This research was supported by the Advanced Research Projects Agency of the Department of Defense, under contracts F30602-70-C-0300, monitored by Rome Air Development Center, Griffiss Air Force Base, New York 13440, and DAHC15-73-C-0363.

TABLE OF CONTENTS

PREFACE

The apparent power of the Digital Signal Processor as a tool for processing information such as pictures and sound is very promising indeed. The variety, precision, and especially the complexity of practical processing schemes that can be realized using digital techniques far exceeds those previously available.

In this research we strive to demonstrate a level of complexity and sophistication in the digital processing of sound waves, which is exemplary of the promise described above. We hope not only that the specific achievements of the project will be useful in and of themselves, but also that the level of endeavor and degree of success will lead others to the exploitation of similar tools in meaningful and useful ways.

Two technical problems are addressed by this research. The first is that of removing interfering background noises from recorded voice signals. This is a classic problem. Sources of background noise are many and varied. To mention a few we include: electronic background noise, noises introduced by the recording medium, environmental noises such as extraneous mechanical noises, or interfering noises similar to those of interest. The example considered in this research involves the second and last of these.

The second technical problem addressed here is that of providing speech analysis and synthesis methods which provide the very highest quality voice characteristics. As they have been known traditionally,

speech analysis-synthesis systems have been developed with a primary emphasis on their ability to reduce bandwidth requirements. This emphasis may have been too great with too little attention having been paid to quality, a characteristic effecting the ultimate user in a nonspecific but strong manner. The very nature of the voice signals in the problem attacked places a very strong emphasis on the issue of quality. The result of this emphasis has been to provide not only a solution to this problem, but some strong insights into the issue of quality which has already influenced separate research projects in our laboratory.

The reader with background in signal processing and electrical communication theory, will recognize the classic difficulty in using traditional thinking to provide a method for separating signals such as speech and accompanying background noises. The problem with the classic approach is that it tends to presuppose that the separation will be effected by linear filtering techniques. This tendency is the result of a long standing tradition and practical compulsion for using linear means in filtering electrical communication signals. This compulsion is primarily a practical one stemming from the availability and relative low cost, as well as the theoretical understanding of the working of such systems. As is well known, when such systems are employed, it is impossible to separate signals which are characterized, in the Fourier sense, by frequency components of the same fundamental period. In the case of complex signals such as speech and noise, an infinity of frequency components are needed to synthesize the waveforms in question. Nevertheless, when these component frequencies are the

same for both the desired signals and the undesired ones, separation remains impossible by linear means. It is common for those not involved regularly with electrical signal processing to forget that this constraint is characteristic only of linear processors and to assume that it is a more general limitation which must always be confronted. Such is not the case. As a matter of fact, the processes used in the research reported here are specifically not linear. Were it not so, the theory tells us that the project would have been unsuccessful. On the other hand, without the digital signal processor as a vehicle, the ideas involved, as successful as they may be, would not be able to be considered on a practical basis with present or short term available technology.

Dr. Thomas G. Stockham, Jr.
July 1973

ABSTRACT*

This report describes research into the problem of rectification of sound recordings made under adverse conditions and communicated and recorded with a great deal of noise. In the course of this research, a number of refinements have been made to the process of digital speech synthesis through new vocoder techniques.

The particular case under investigation was that of old, noisy recordings of a singing voice and the immediate goal was the separation of that voice from wide band noise, orchestral accompanyment and recording noise. This has been accomplished by the development and refinement of the homomorphic vocoder as a filtering device.

---

*This report reproduces a dissertation of the same title submitted to the Department of Electrical Engineering, University of Utah, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

I. INTRODUCTION

This research represents the development of a system for the separation of singing voice from musical accompaniment and wide band surface scratch as found on early acoustic disc recordings. The separation system employs a resynthesis of the singing voice using a mechanism, the homomorphic vocoder, based on the properties of voice signals.

The homomorphic vocoder separately estimates the contributions of vocal tract excitation and vocal tract articulation. When the homomorphic vocoder is applied to voice signals corrupted by noise, the corruptions affect the estimated excitation and articulation components. However, these components have been found to be regular over long time intervals, while the corruptions are not. Therefore, the effect of the corruptions can be attenuated by constraining the synthesized excitation and articulation functions so that they are also regular over long time intervals.

Further improvement in the quality and naturalness of the synthetic voice signal can be made by implementing several additional procedures. The mechanical recording system used to produce the recordings was incapable of responding to high frequencies, therefore, an attempt was made to synthesize artificial high frequencies. The voice signals occurring at the beginnings and endings of singing segments often have much lower energy than the accompanying noise. This makes accurate estimation of the actual vocal tract excitation and articulation functions

difficult. This difficulty can be overcome by substituting estimations from adjacent, high signal energy sections for the unrecoverable sections. The synthetic signal does not have the benefit of room acoustic dispersion, consequently an attempt has been made to introduce artificial room reverberation.

The major contribution of this research is the presentation and implimentation of the notion of filtering a voice signal from noise using an analysis-synthesis system. Although many of the techniques used in this implimentation are derived from the literature, several have been developed exclusively for this task. The notion of pitch synchronous reverberation has not been presented elsewhere. This procedure, which interrelates neighboring impulse response estimates, is responsible for much of the success of the filtering process. The heuristic and interactive pitch correction processes, though based on classical notions, have been refined and adapted for this task. Finally, the notion of extrapolating the synthetic signal into intervals with low signal to noise ratios by using information derived from intervals with high signal to noise ratios has been developed successfully in this application. These contributions represent steps not only in the solution of the problem of filtering voice signals from noise, but lead to possible solutions of problems in many other areas.

A successful scheme for extracting a signal from a background of noise could be applied to a variety of communication problems. The homomorphic vocoder has been successfully used in bandwidth reduction, helium speech correction, and expansion-compression systems [1],[2],[3]. This work explores its utility as a nonlinear filtering mechanism for voice signals and introduces several modifications to further adapt the vocoder to this application. This modified vocoder might be applied to extraction of sonar reverberations from biomarine sounds, reconstruction of voice messages overlaid with vibrational noise or to any problem where the parameters of a biologically produced sound need to be estimated and recreated free of other accompanying mechanical noises.

II. PRINCIPLES OF THE HOMOMORPHIC VOCODER

2.1 Historical Background

In 1939 Dudley developed a method of coding voice signals which significantly reduced the bandwidth necessary for voice transmission [4]. This method, called the channel vocoder, functions as follows: First, the voice signal is passed to a bank of bandpass filters before direct transmission. The outputs of these filters is then rectified and smoothed, producing channel signals, which are transmitted together with a signal containing voiced/unvoiced/pitch information. An input signal generated from the transmitted voiced/unvoiced pitch signal excites a bank of bandpass filters identical to those at the sending end. The output signals of the bandpass filters are modulated by the channel signals. The input signal is generated as white noise for unvoiced excitation and a sequence of impulses at the pitch rate for voiced excitation.

Not only was Dudley able to produce intelligible speech at one sixth the bandwidth required to send the actual voice signal, but his channel vocoder serves as a basis for the vocoders presently in use for speech synthesis. The success of Dudley's process not only provided a solution for the bandwidth reduction problem, but also demonstrated some important perception properties of voice signals. Utilizing this process, intelligible speech could be generated by synthesizing a signal that preserved both the excitation function of the original voice signal and the short time spectral characteristics or articulation function.

The numerous improvements in vocoders made since 1939 primarily resulted in more efficient and accurate estimations of the excitation function [5], [6], and in better techniques for estimating and coding articulation functions [7], [8]. The channel vocoder's short time spectrum is formed from the output of a bank of bandpass filters, which is sampling along the frequency axis at fixed frequencies. Subsequent workers, modelling the vocal tract as a resonant chamber, postulated the existence of a small number of resonant frequencies called formants which would describe the short time spectrum. The short time spectrum was reconstructed from its most prominent four or five formants and their associated Q's. This system, known as the formant vocoder, successfully utilized the fact that the formant frequencies change slowly and can be tracked over long intervals of time [8].

In 1966 Oppenheim presented an analysis of the vocoder based on his work with homomorphic transformations. This analysis became known as the homomorphic vocoder. Oppenheim reasoned that the vocal tract was essentially a stationary linear system over short intervals of time. The output of such a system is known to be the convolutional combination of the excitation function and the impulse response of the system. Oppenheim's previous work had shown that when two functions are combined by convolution, they obey transformational rules equivalent to addition and scalar multiplication. When this condition applies, there exists an invertible, non-linear, homomorphic transformation which maps these functions into new functions combined by ordinary addition and scalar multiplication. Linear filters can be applied to these after they

have been transformed into a space where they are combined by ordinary addition [9].

An application of the Fourier transform maps functions combined by convolution into functions combined by multiplication. If the logarithm of the result of this transformation is computed, then the functions combined by multiplication can be further transformed into functions combined by ordinary addition. Thus, the log spectrum of a section of voice signal contains components which are derived from the excitation function and from the impulse response of the vocal tract, combined by ordinary addition. Oppenheim's homomorphic vocoder applied linear filtering of the log spectrum to effect a separation of the excitation function and the impulse response of the vocal tract, yielding some of the highest quality speech obtained by vocoder processing.

2.2 The Convolutional Model and Discrete Representation

The homomorphic vocoder models a singing voice signal as the output of a piecewise time invariant linear system [9] . Probably the most important characteristic of time invariant linear systems is that each of them is completely characterized by its impulse response. For a system with impulse response h(t), its response to an input f(t) is

$$g(t) = \int_{\infty}^{\infty} f(T).h(t - T).dT = f(t) \circledast h(t) \qquad (1)$$

that is, g(t), the output of the system is the convolution of f(t) with h(t) [10]. Therefore, to regenerate a singing signal without reproducing the background noise, one must only determine the singing

voice's excitation function and the sequence of impulse responses that are associated with it for each time interval.

The use of the convolution integral (1) assumes that the input signal is generally defined and continuous for all t in the interval $-\infty$ to $+\infty$. The impulse response, h(t), and the output of the system, g(t), are similarly defined. The convolution integral can be simplified by restricting the input function and impulse response. Bandlimiting the functions f(t) and h(t) aid in simplifying the convolution integral. Signals whose component functions are so limited can be processed as discrete samples and then regenerated, producing the equivalent continuous counterparts. When f(t) and h(t) are bandlimited and sampled at a rate exceeding the Nyquist rate, then, as proven by the Sampling Theorem, the signals can be completely recovered from their discrete samples [11]. Low pass filtering of the noisy recording effectively band limits the functions f(t) and h(t) and the noise signal, since the original signal contains all of these components. For this analysis, we have lowpass filtered the noisy signal at 4 K Hz and sampled at 10 K Hz. A grace band of 1 K Hz was allowed since the analog filters used cannot attenuate completely and immediately above 4 K Hz.

The convolution integral (1) of this filtered, bandlimited signal can now be replaced, without introducing any approximations, by the sum

$$g(K) = \sum_{J=-\infty}^{\infty} f(J).h(K - J) \qquad (2)$$

where g , is the Kth sample of the output function, f(J) is the Jth sample of the input function, and h(K-J) is the (K-J)th sample of the impulse response.

Restricting the impulse response of the vocal tract, h(t), to be non-negligible only over a finite interval further simplifies the convolution integral. The vocal tract is a mechanical system and is realizable. Thus, its impulse response is zero for all t less than zero. The vocal tract also has both internal and external viscous frictional forces acting upon it, thus, its impulse response tends to zero for increased t. A sequence of equally spaced samples from a function which is non zero only over a defined interval is itself of finite length. Therefore, for finite length impulse responses, the sum in (2) becomes

$$g(K) = \sum_{J = K - M}^{K + M} f(J).h(K - J) \qquad (3)$$

whenever $h(L) \neq 0$ for $-M \leq L \leq M$, and

$= 0$ otherwise

This restriction reduces the convolution (1) to a sum which can be computed in a defined number of operations.

Within this discrete framework, the task of generating the samples of a singing signal without reproducing the background noise reduces to determining the excitation function and processing the sequence of samples associated with the impulse responses.

III. EXCITATION FUNCTION DETERMINATION

Oppenheim's formulation of the homomorphic vocoder generates the speech excitation function from measurements of pitch rate and a voiced-unvoiced decision [9]. The generated function for voiced segments consists of a sequence of unit impulses or unit samples occurring at intervals corresponding to the pitch periods. The unvoiced segments are generated as a waveform with a flat spectrum, for example, a train of impulses with random polarity. In addition to voiced and unvoiced components, recorded singing voice signals have many intervals which contain either silence or only background accompaniment. The vocal tract excitation function employed in this paper consists of sections which are voiced, unvoiced, and silent.

Lungs and glottis, the physiomechanical structures which produce the vocal tract excitation function, are limited in the rate at which they can change positions. Therefore, the estimated excitation function for singing voice has been constrained to remain constant over each successive 6.4 millisecond interval. This is not a severe restriction since even the highest pitched samples will contain no more than three pitch periods during this brief interval. The calculation of the input function for 6.4 millisecond intervals rather than for each sample also results in considerable computational savings.

The process which has been implemented to determine the excitation function over all successive 6.4 millisecond intervals in a selection has several stages. First, it is assumed that all intervals are voiced,

then an estimate of the pitch rate over each interval is made by cepstral pitch detection [6]. Finally, this sequence of pitch rate estimates is segmented as to voiced, unvoiced, or silent excitations by utilizing a heuristic program and manual interaction.

3.1 Cepstral Pitch Detection Procedure

The computation of the cepstrum is made using a modification of Noll's procedure [12]. The input signal is lowpass filtered at 4 K Hz. and sampled at 10 K Hz. to 14 bits to permit digital representation. For each estimate, a Hanning window of 256 samples is applied. This data window width is a compromise between shorter windows, which degrade the accuracy of the pitch rate estimates, and longer windows, which introduce averaging errors due to incorporation of pitch rates from neighboring intervals. Zero samples are appended to the sequence of windowed samples, a 512 point discrete Fourier transform (DFT) is performed, and the log magnitude of the spectrum is computed; the zeros appended prior to transforming provide a two fold interpolation in the frequency domain and reduce the effect of aliasing in the cepstrum. A 512 point Hanning window is applied to the interpolated log magnitude spectrum, 1536 zeros are symmetrically concatenated, and a 2048 point inverse DFT is computed yielding the cepstrum. This application of the Hanning window distorts the cepstrum less than simpler windows, even though the high frequency components in the log magnitude spectrum are attenuated. The concatenation of zeros results in an interpolation of the cepstrum to a resolution of 0.025 milliseconds permitting a more exact estimation of pitch period. The estimate of pitch period is made

from the quefrency with the maximum cepstral value over an interval from 1.65 milliseconds to 12.0 milliseconds. The pitch rate estimate is the inverse of this quefrency.

A block diagram of this process is presented in Figure 1, with typical intermediate waveforms represented in Figures 2, 3, 4, 5, 6, 7, and 8. The results of this pitch detection scheme over 1024, 6.4 millisecond intervals are shown in Figure 9, where the pitch rate in Hertz is plotted against time.

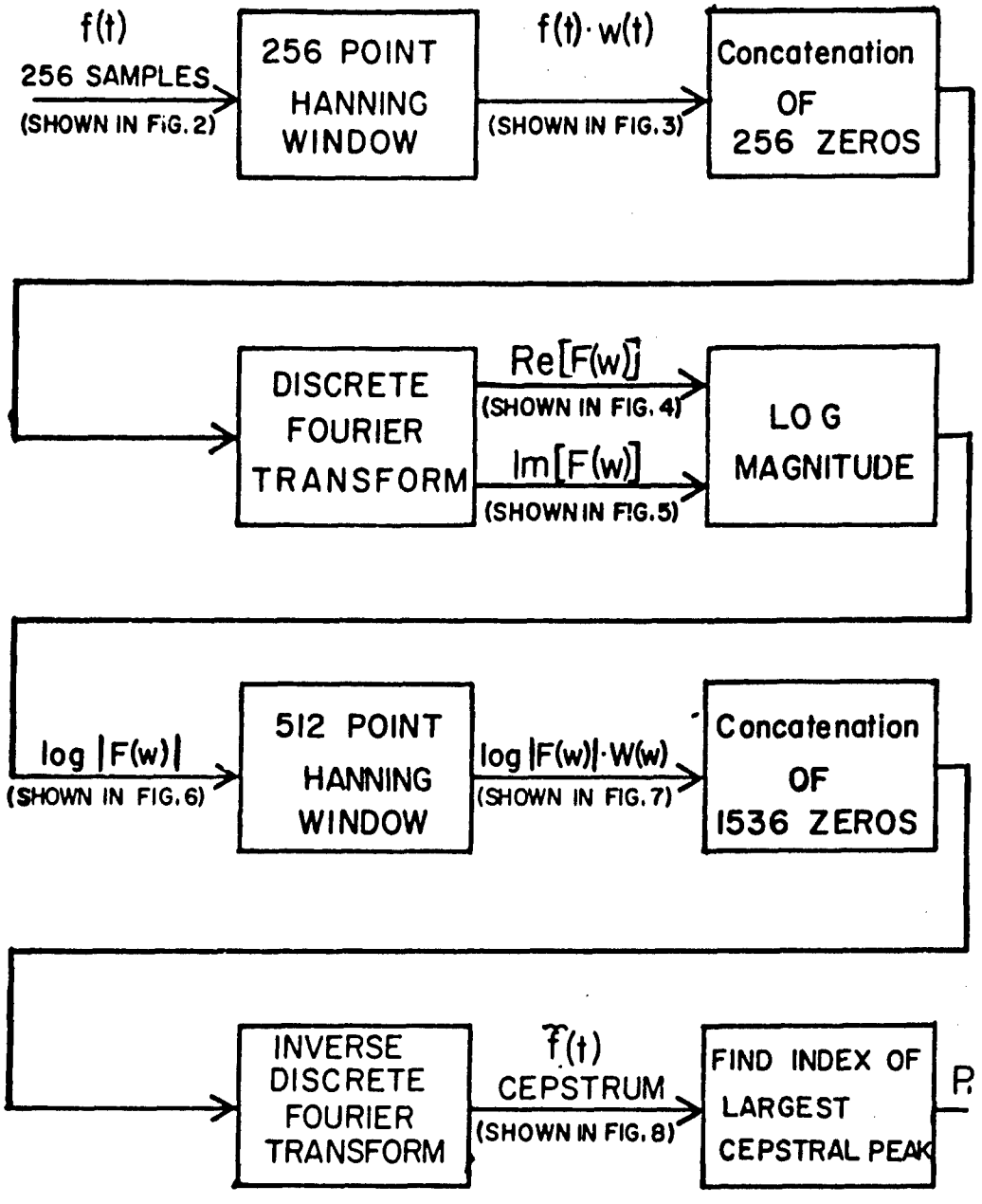

FIGURE 1. BLOCK DIAGRAM OF CEPSTRAL PITCH DETECTION PROCESS.

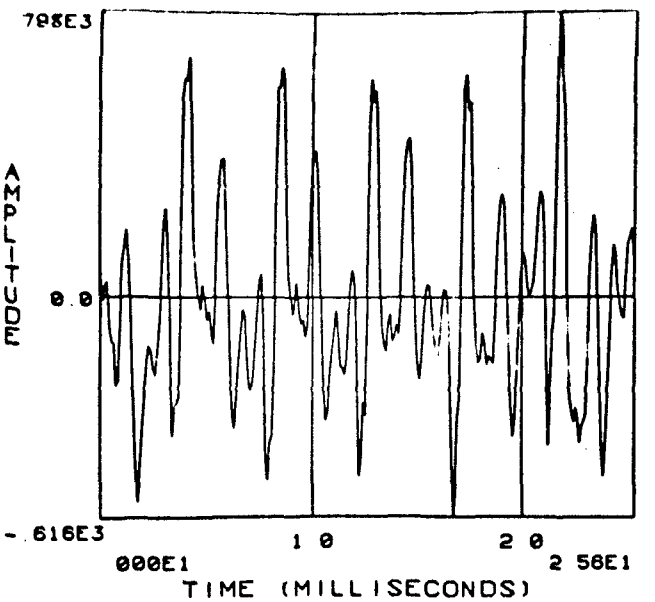

Figure 2. A 25.6 millisecond segment of the noisy signal. This is over the ā sound in "Recitar!".

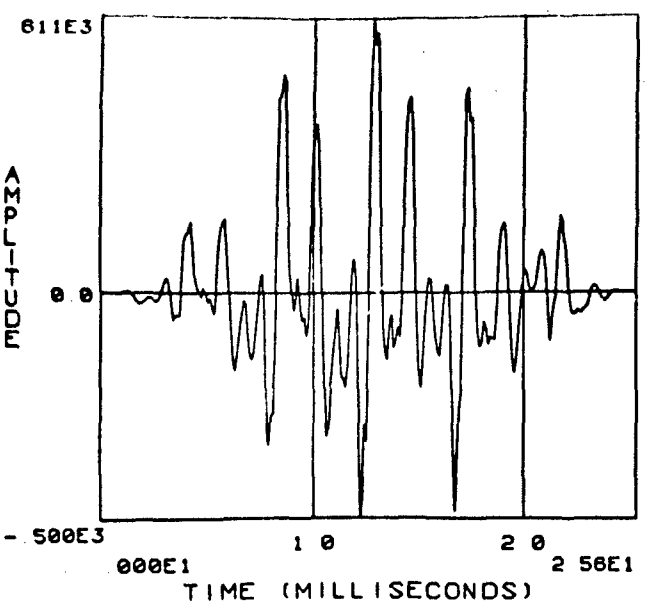

Figure 3. A 25.6 millisecond segment windowed with a Hanning window of width 25.6 milliseconds.

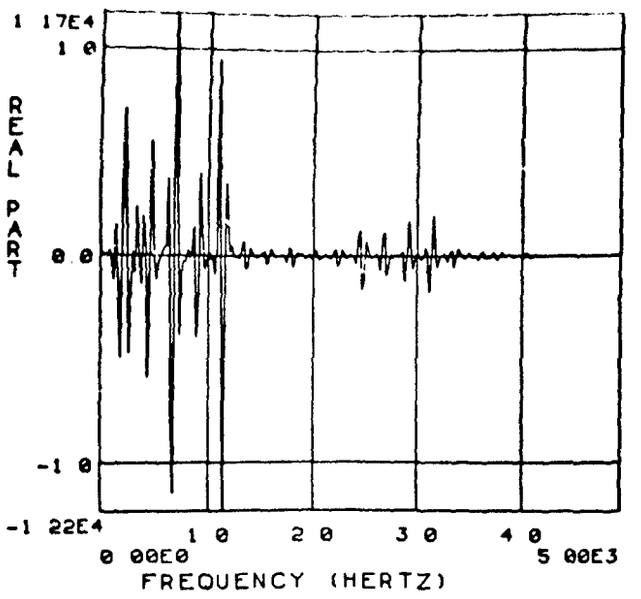

Figure 4. The real part of the estimate of the spectrum of a 25.6 millisecond section of singing.

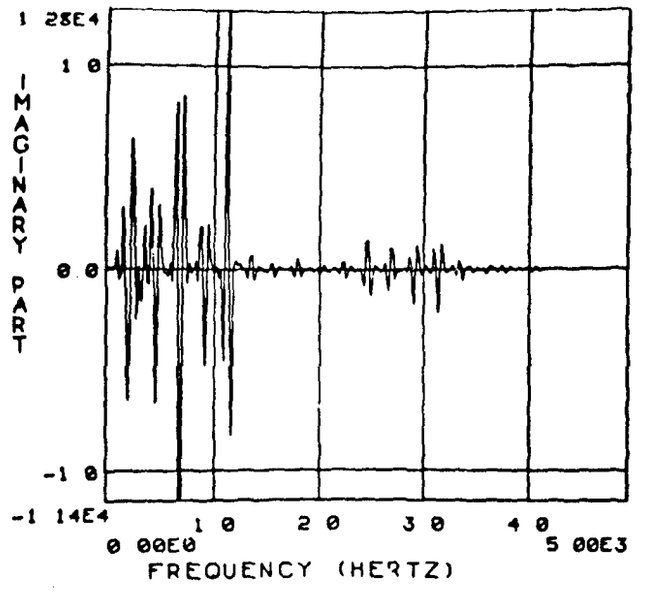

Figure 5. The imaginary part of the estimate of the spectrum of a 25.6 millisecond segment of singing.

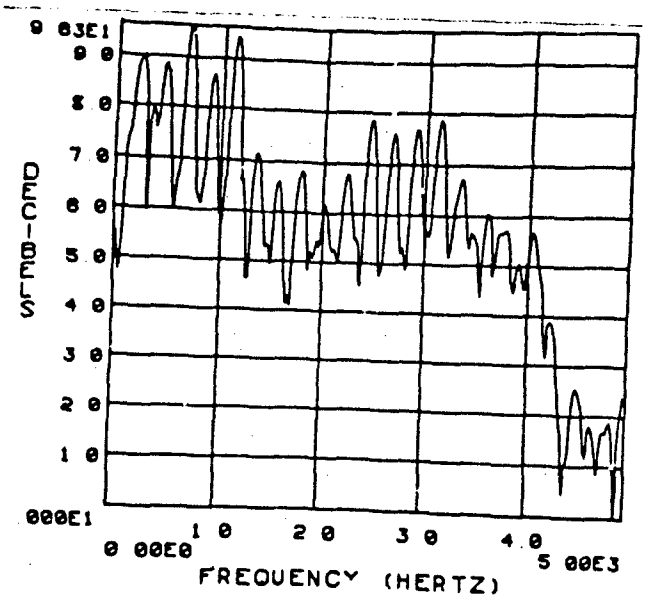

Figure 6. The Logarithm of the Magnitude of the Spectrum of a 25.6 millisecond interval of singing.

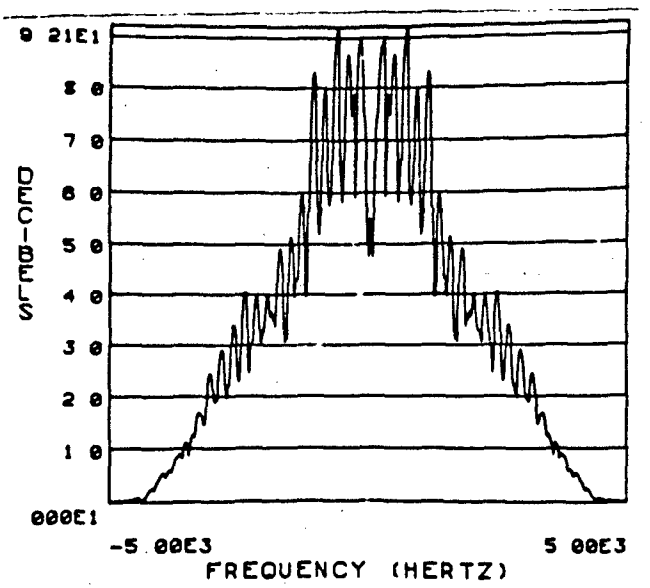

Figure 7. The Hanning windowed two-sided log magnitude Spectrum of a 25.6 millisecond interval of singing.

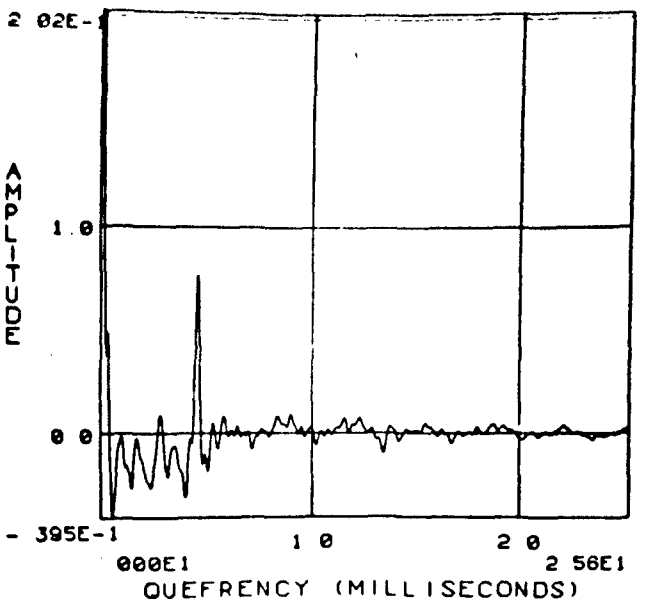

Figure 8. The Cepstrum of a 25.6 millisecond interval of singing. (Excludes first 5 quefrencies).

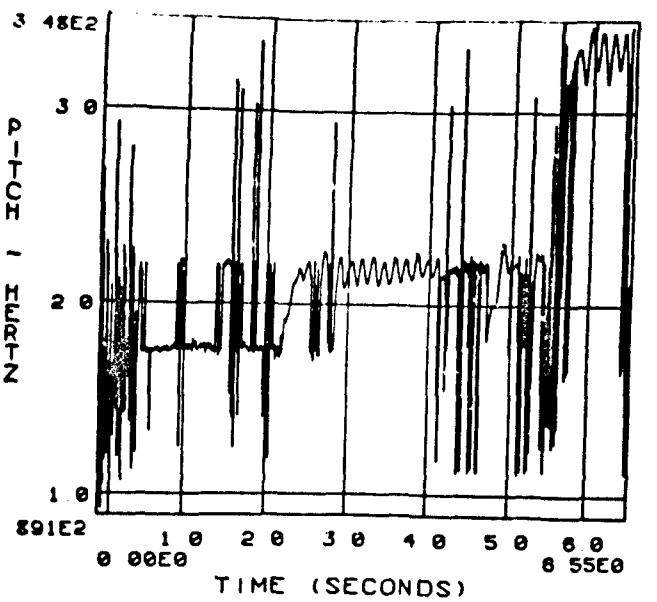

Figure 9. The Raw Cepstral Pitch Estimates for 1024 6.4 millisecond intervals (6.55 seconds).

3.2 Heuristic Segmentation Procedure

Once the pitch estimates have been computed, they must be segmented into voice, unvoiced, and silent sections. Segmentation can be accomplished by listening to the original noisy signal, screening the corresponding section, and then manually classifying each pitch estimate. This method is not only tedious, but also fails to utilize an important characteristic of the pitch rates: the pitch rate for each 6.4 millisecond interval within a voiced segment closely approximates the pitch rates of its adjacent intervals. A two pass heuristic procedure which utilizes this property has been implemented. This automatic classification procedure identifies many of the voiced sections of the signal, thus reducing the number of pitch estimates that must be manually classified.

During the first pass of the heuristic program, regression lines are calculated which approximate the pitch estimates for each group of five consecutive 6.4 millisecond intervals. The pitch estimates for these five consecutive intervals are marked as voiced if the residual variance from their regression line is less than a threshold value of 20 $Hz.^2$. This threshold value was chosen after observing that the pitch rate variances were on the order of 10^4 $Hz.^2$ over manually identified unvoiced and silent sections, but less than $1.0Hz.^2$ over voiced sections. Figures 10 and 11 illustrate data before and after this first pass of the program.

The second pass classifies as voiced any segment consisting of less than three 6.4 millisecond intervals which separate two sections marked as voiced. Singing voice signals cannot be unvoiced or silent over

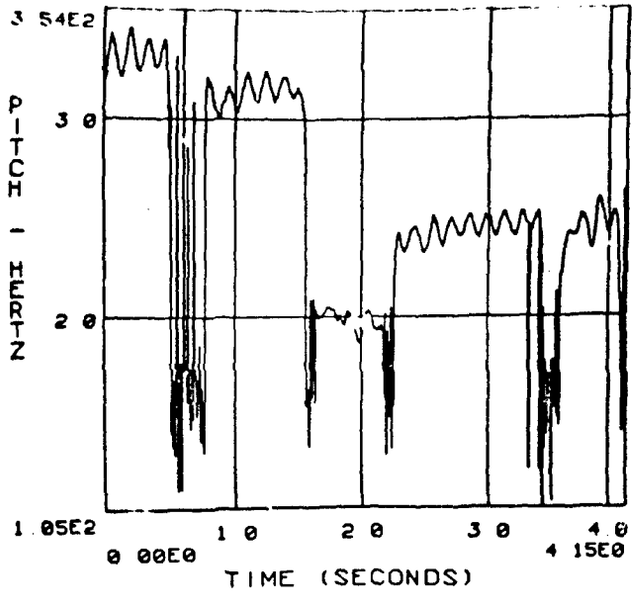

Figure 10. A 4.15 second section of raw pitch estimates before application of the first pass of the heuristic segmentation program.

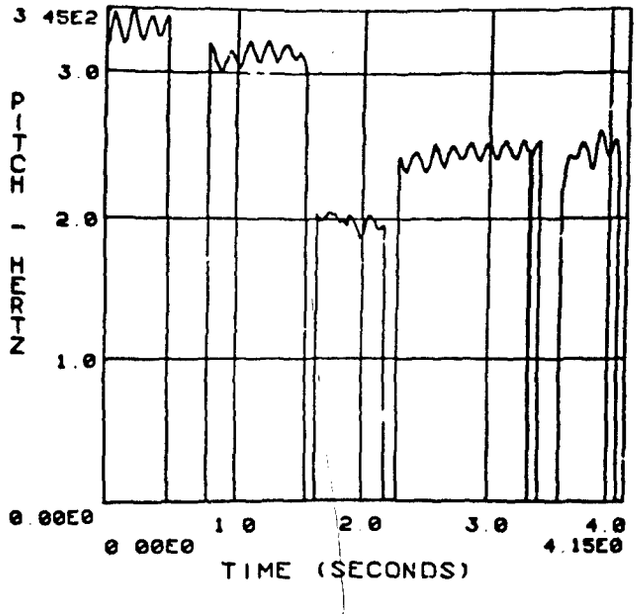

Figure 11. The 4.15 second section of raw pitch estimates shown in Figure 10 after application of the first pass of the heuristic segmentation program.

such short sections, therefore, misclassification of these intervals is due to errors in the original rate estimation. New pitch rates are computed for these short intervals by linear interpolation from the endpoints of their adjacent voiced sections. An example of the effect of this second pass is shown in Figures 12 and 13.

3.3 Interactive Segmentation Procedure

The automatic program for identifying voiced intervals processes the bulk of the singing voice signal. However, misidentification and omission errors are made by this program and must be corrected by manual interaction. Segments consisting only of musical accompaniment are sometimes classified as voiced. If the pitch estimates marked voiced by the heuristic program are graphically displayed, these misidentification errors are easily recognized. As Figures 14 and 15 illustrate, the variation of voiced pitch profiles is small. Vibrato is present only over extremely long intervals. In contrast, the profiles of pitch estimates made from intervals containing only musical accompaniment display a constant variation. These results demonstrate that musical accompaniment is a polyphonic sound with a unique pitch over those short intervals dominated by one part of the accompaniment.

Omission errors are most frequent for intervals at the start and finish of voiced segments, when the signal contains singing signal, background accompaniment and surface scratch of approximately equal intensity. The pitch estimates for such intervals commonly contain errors. The heuristic program fails to mark these fallacious estimates as voiced because they do not fulfill the voiced decision requirements

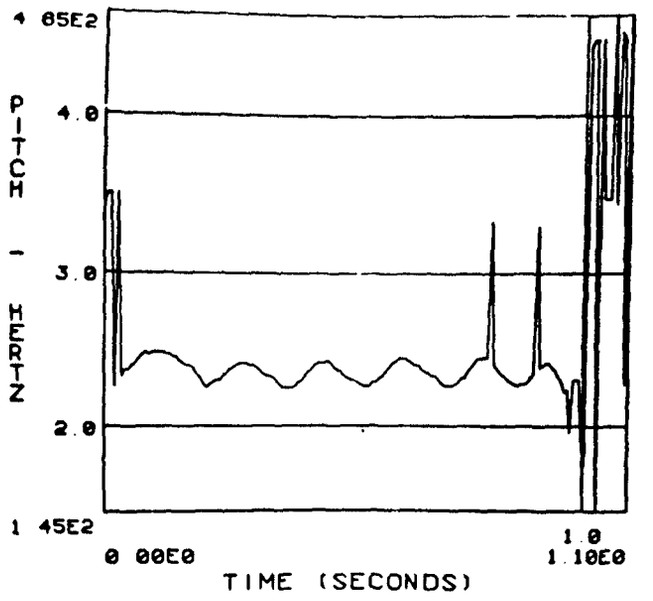

Figure 12. 1.1 seconds of raw pitch estimates showing pitch detection errors.

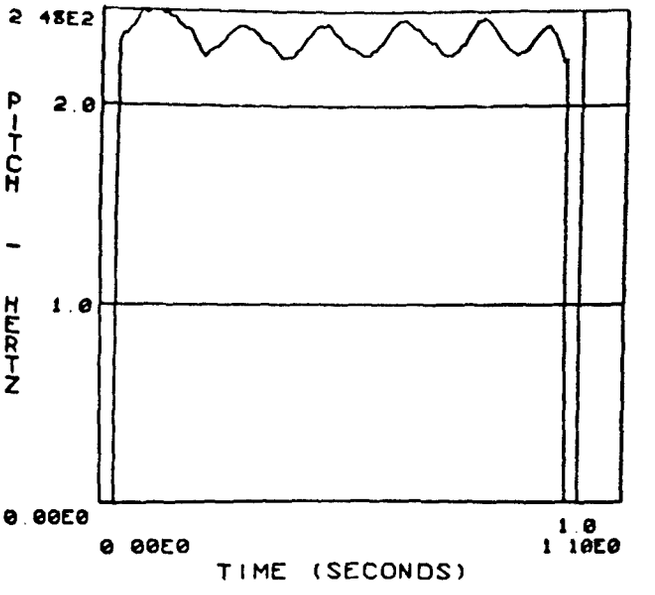

Figure 13. 1.1 seconds of pitch estimates after the applications of the first and second pass of the heuristic segmentation program.

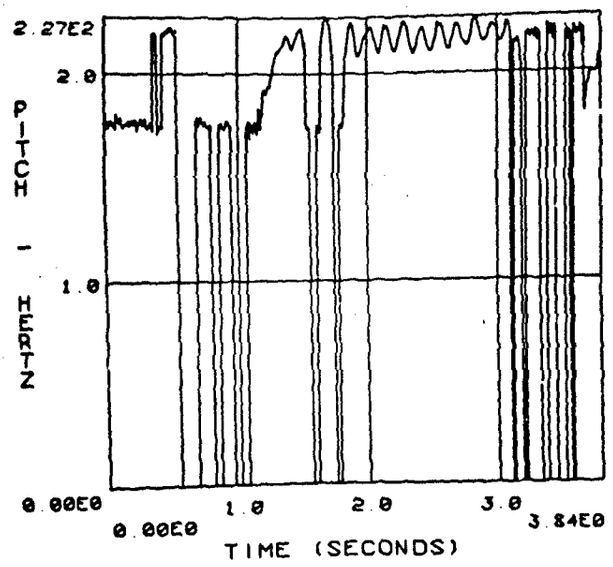

Figure 14. Pitch estimates segmented by the heuristic program but having orchestral segments identified as voiced.

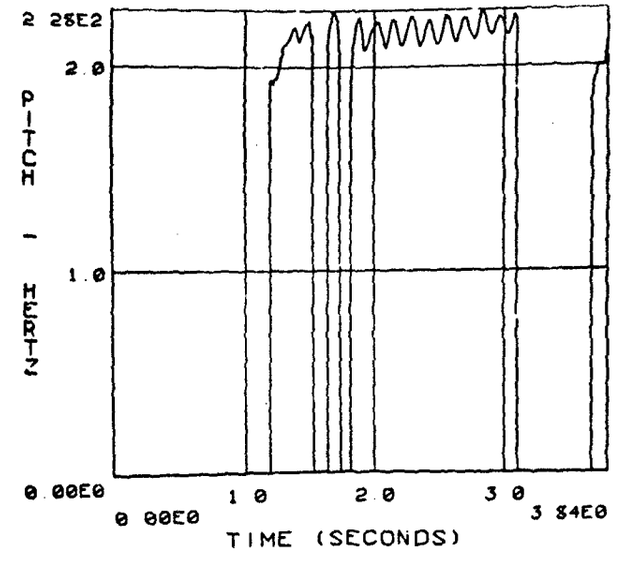

Figure 15. The estimates of Figure 14, manually corrected to eliminate pitch due to accompaniment.

of the program. A graphical display is used to view the pitch estimates. These erroneous estimates often become obvious by their large deviation from the homogeneous pitch profiles of the correct estimates. Figures 16 and 17 are examples of the pitch profiles of this type of error before and after correction. The incorrect estimates are manually marked for interpolation. Before each synthetic recording is made, all such noted intervals are linearly interpolated along a line connecting the endpoints of adjacent voiced sections. These new pitch estimates are included in the synthesis process.

Interactive identification has also been the only successful technique found for designating unvoiced intervals and distinguishing them from surface scratched silent sections. The unvoiced sections must be identified by auditioning the original recording. The excitation function for the unvoiced sections are taken to be samples of equal magnitude with polarity determined by a pseudo-random number generator, a procedure used by Oppenheim and equivalent to the procedure used by Dudley [3], [4]. Any interval not marked as voiced or unvoiced is designated silent and its excitation function is taken to be zero. After a synthetic recording has been made, it is auditioned and any remaining omissions or misidentifications are noted and manually corrected for the next synthesis.

3.4 Evaluation

The success of the excitation determination process used in this analysis can be attributed to its binary nature. The processes initiates a sequence of yes or no decisions such as choosing between singing

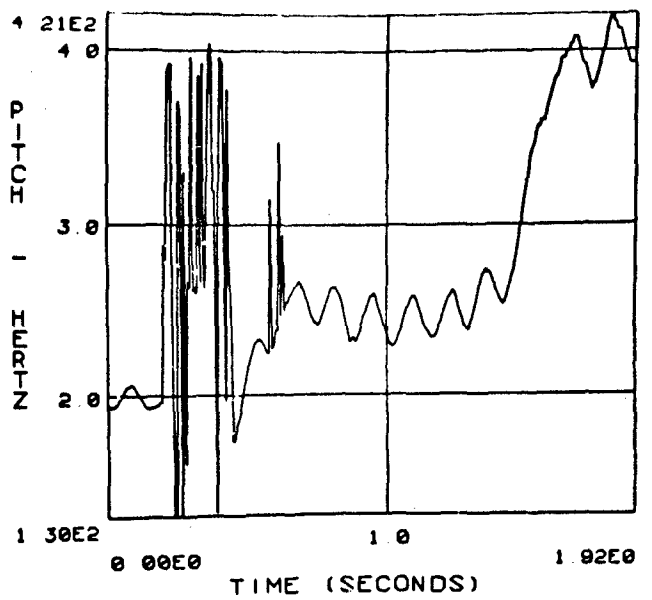

Figure 16. Raw pitch estimates showing pitch detection errors not corrected by the heuristic segmentation program.

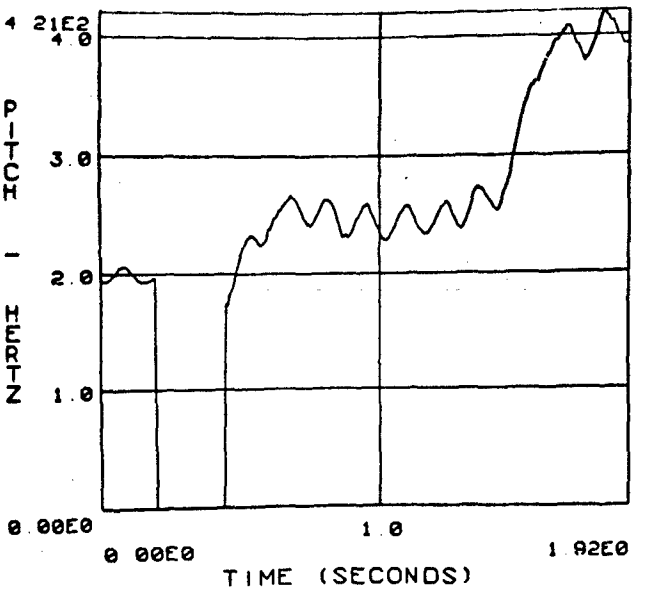

Figure 17. The pitch estimates of Figure 16 having been processed by the heuristic segmentation program and manually corrected.

voice or nonsinging, and then for the singing voice sections, voiced or nonvoiced segments. The graphical and auditory feedback parts of the process also contribute to the overall success of the excitation function determination.

Extreme care was required during determination of the excitation function. This determination not only controls the quality of the voiced segments by its regulation of pitch, but also distinguishes between sections which are silent and should be generated as zeros. Any error in the excitation function becomes explicitly noticeable in the synthetic signal.

The major failure of this process is the abruptness created by the beginnings and endings of the singing segments. After listening to the synthetic recordings, it becomes clear that the singer must have been more gradually increasing and decreasing the volume of his voice in the original recording. However, the voice signal, including its pitch, is deeply imbedded in the noise of these starting and finishing segments. A partial solution to this problem, presented in the chapter on synthesis, uses the first and last detectable pitch and impulse responses as best approximations for the pitch rates and impulse responses for these beginnings and ending sections.

IV. IMPULSE RESPONSE ESTIMATION

Cepstral liftering as applied in the homomorphic vocoder [13], formant detection [8], and predictive coding [14] have all been successfully used for estimation of the vocal tract impulse response from noise free signals. Estimation of the impulse responses for acoustically recorded singing signals is complicated by a lack of high frequency information and the competing presence of musical accompaniment and noise signals.

The cepstral liftering technique, with modifications, provided the most direct means of generating high quality impulse responses with minimum, maximum, and zero phase for the singing voice signal. In addition to processing the singing signal by cepstral liftering, a first order approximation to the missing high frequencies is introduced into the spectrum of each impulse response. Constraints are also implemented to compensate for the presence of corrupting noise in the original signal.

4.1 The Short Pass Liftering Process

The first step in the cepstral liftering process is the estimation of the log magnitude spectrum. A sequence of time domain samples is windowed with a Hanning window and a number of zeros equal to the window width are concatenated. The discrete Fourier transform is applied to this sample yielding a complex spectrum of the noisy data from which the log magnitude spectrum is computed. Application of a Hanning window

prevents leakage in the estimated spectrum [6] and the concatenated zeros provide an interpolated spectrum and reduce the effect of aliasing in the cepstrum [15]. After computing the log magnitude spectrum, a data window is applied and an inverse DFT performed to produce the cepstrum. Although a Hanning window was applied to the log magnitude spectrum for the excitation function determination in the last chapter, this window is not used here. We have found that use of the Hanning window alternates high frequencies. Since the two sided symmetric log magnitude spectrum is already bell shaped, a Fourier window can be applied. Although this type of window allows considerable leakage, it least attenuates the high frequency spectral values. The use of a one sided window could produce cepstral distortion due to a sharp discontinuity at zero hertz. This use of the two sided log spectrum is consistent with the experiments of Schroeder and Noll [16] concerning log power spectrum windowing for cepstral computation.

When the cepstrum is short pass liftered, the high quefrency values are zeroed. This processed cepstrum determines a zero phase impulse response. If the cepstral values with negative quefrencies are zeroed and those with quefrencies greater than zero are doubled, the result is a minimum phase impulse response. Reversing the roles of negative and positive quefrencies, the same computation produces a maximum phase impulse response [9]. After the phase has been determined, the cepstrum is reduced to the length of the time domain data by symmetrically eliminating the highest quefrencies.

A DFT is performed on this cepstrum yielding a smoothed log magnitude spectrum and phase. The log magnitude spectrum, when exponentiated,

results in the magnitude spectrum, which is used with the phase to produce the complex spectrum of the impulse response. To eliminate the low frequency characteristics of the impulse response, the low frequency coefficients of the complex spectrum are zeroed. An inverse DFT is applied resulting in the estimate of the impulse response.

A block diagram of this process is shown in Figure 18 and intermediate waveforms are shown in Figures 19 through 32.

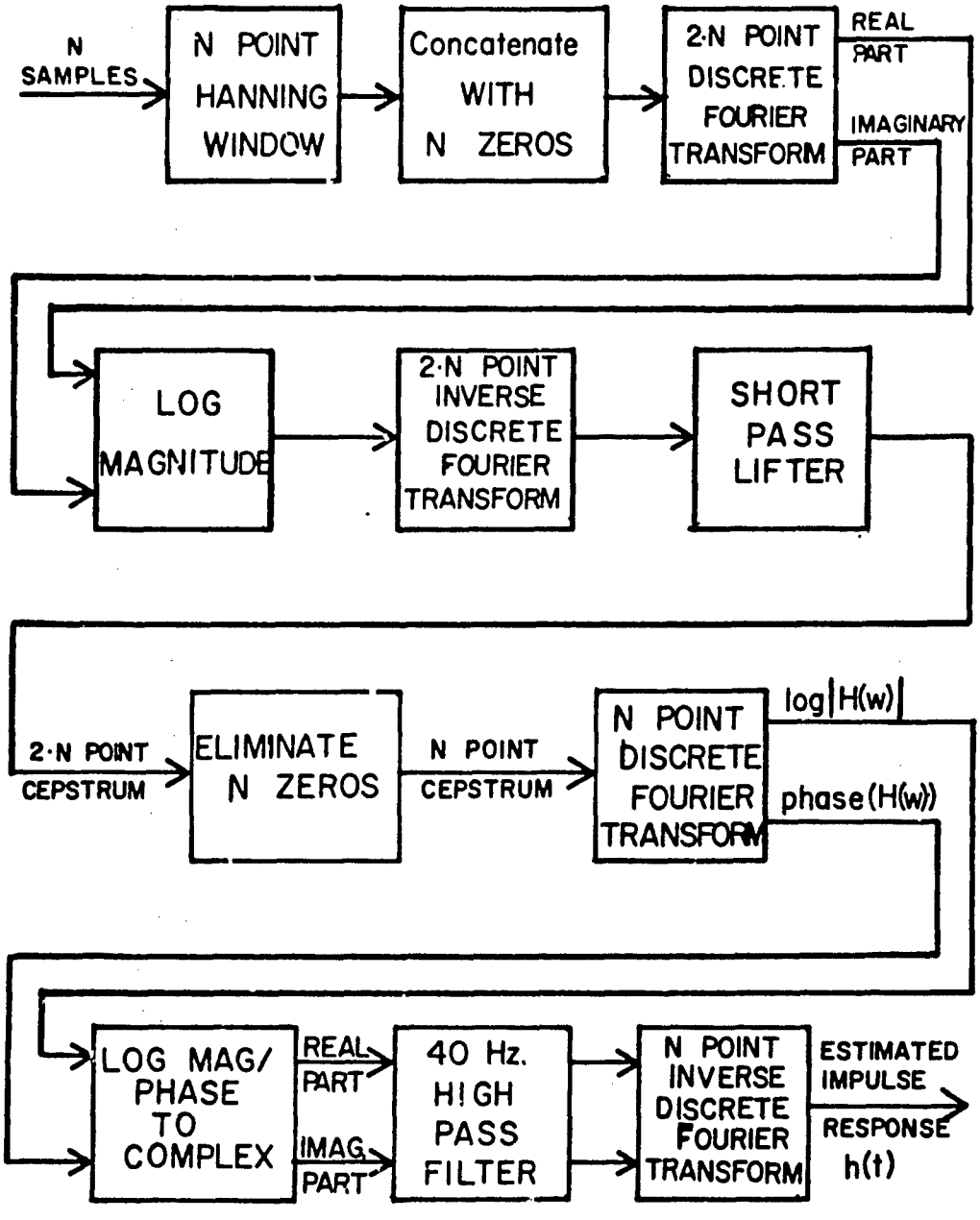

FIGURE 18. BLOCK DIAGRAM OF IMPULSE RESPONSE ESTIMATION PROCESS.

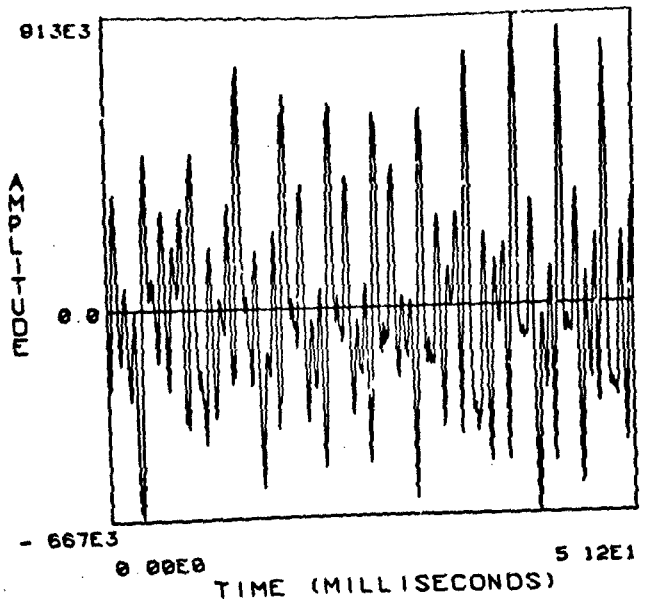

Figure 19. A 51.2 millisecord segment of the noisy signal. This is over the æ sound in "Recitar!"

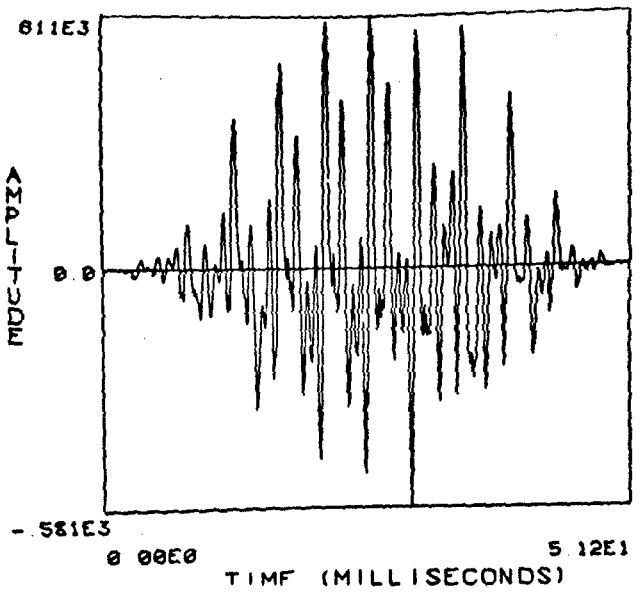

Figure 20. The 51.2 millisecond segment of Figure 19, windowed with a Hanning window of width 51.2 m.s.

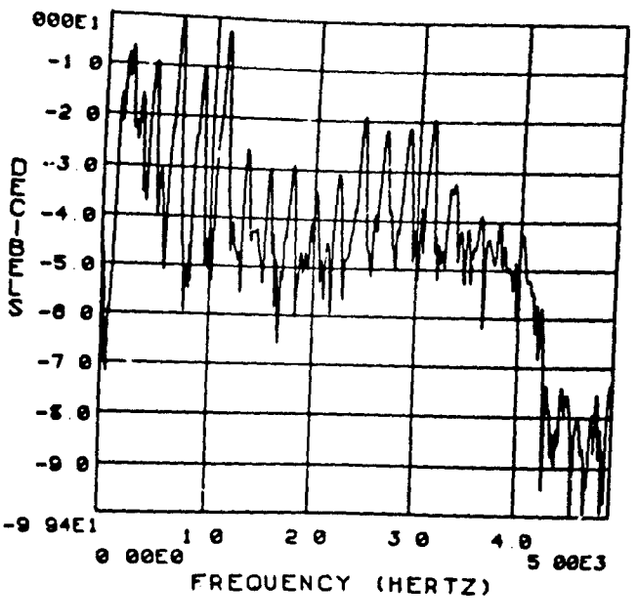

Figure 21. The log magnitude spectrum of the 51.2 millisecond section of Figure 19.

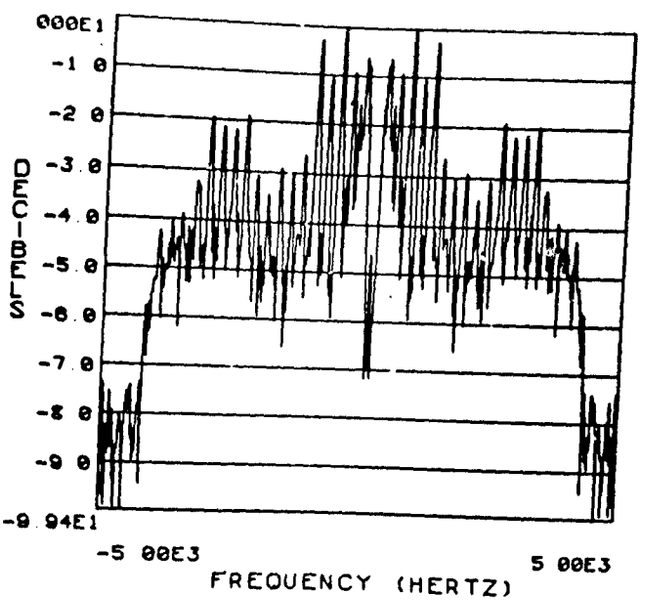

Figure 22. The two sided log magnitude spectrum of the 51.2 millisecond section of Figure 19.

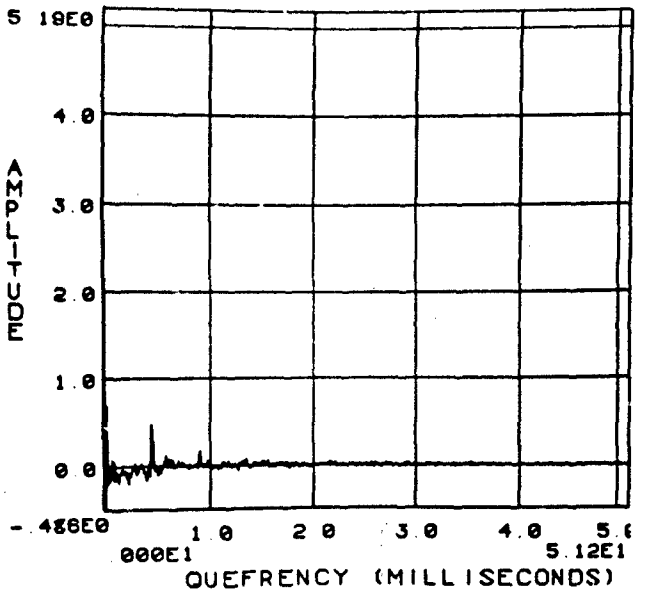

Figure 23. The Cepstrum estimated from the 51.2 millisecond section of Figure 19.

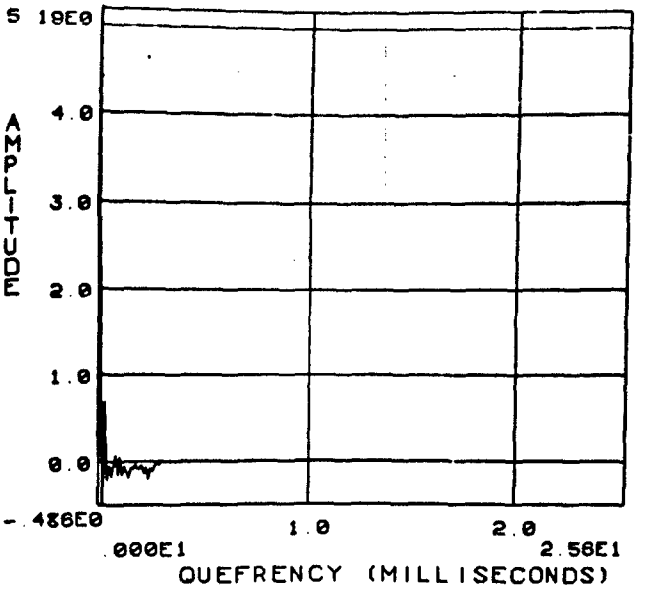

Figure 24. The shortpass liftered Cepstrum from the Cepstrum of Figure 23. Only 30 cepstral values remain.

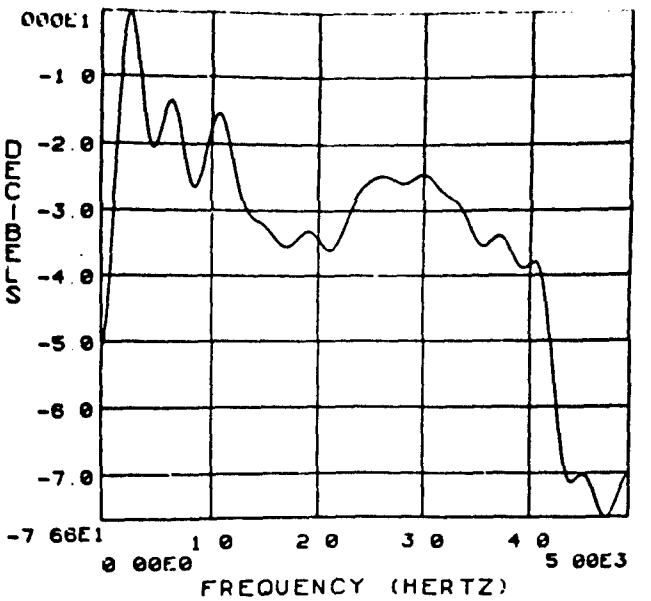

Figure 25. The log maynitude spectrum of the impulse response.

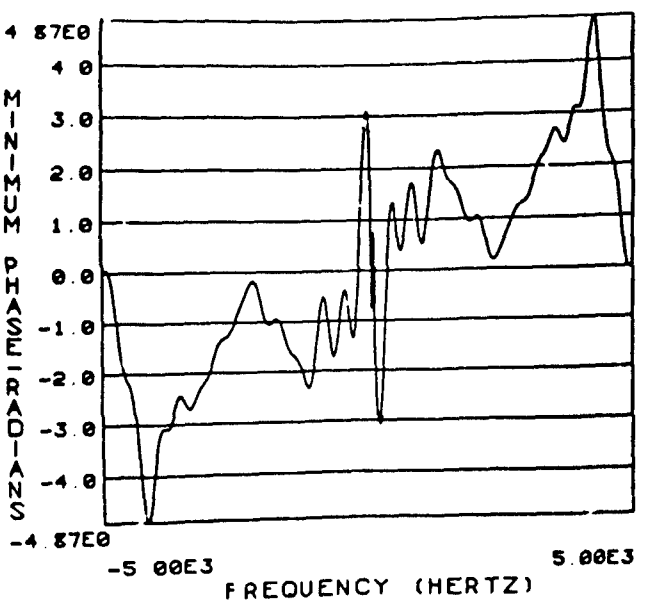

Figure 26. The minimum phase estimate of the impulse response.

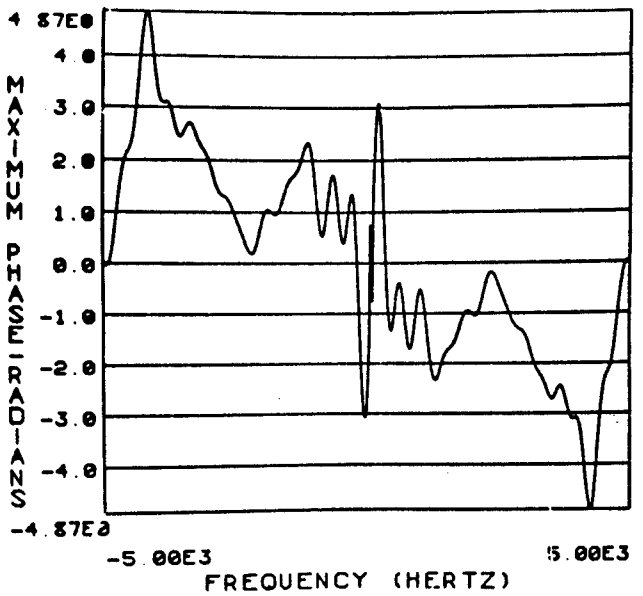

Figure 27. The maximum phase estimate of the impulse response.

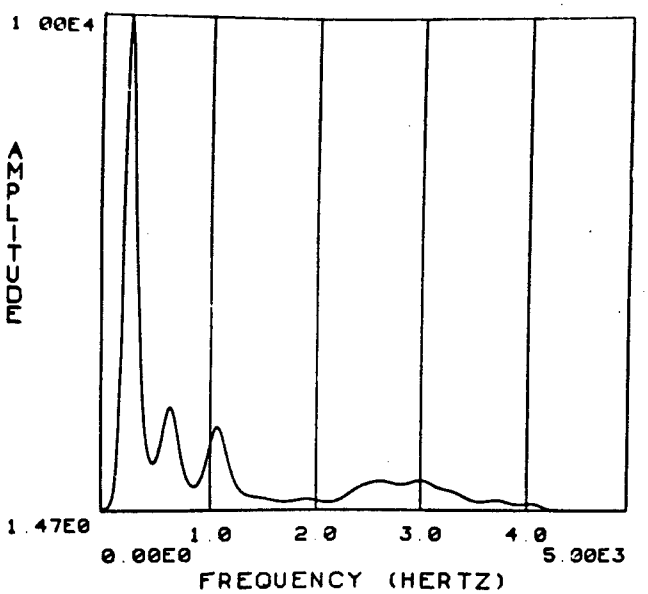

Figure 28. The magnitude of the spectrum of the impulse response.

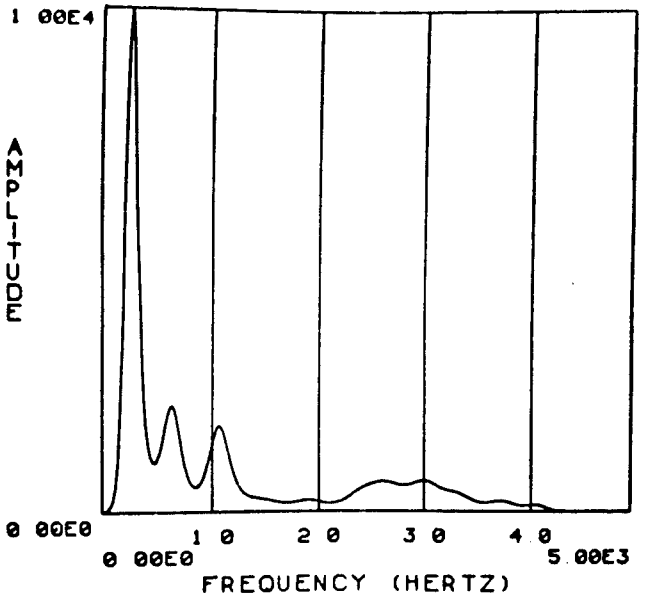

Figure 29. The highpass filtered magnitude spectrum of the impulse response.

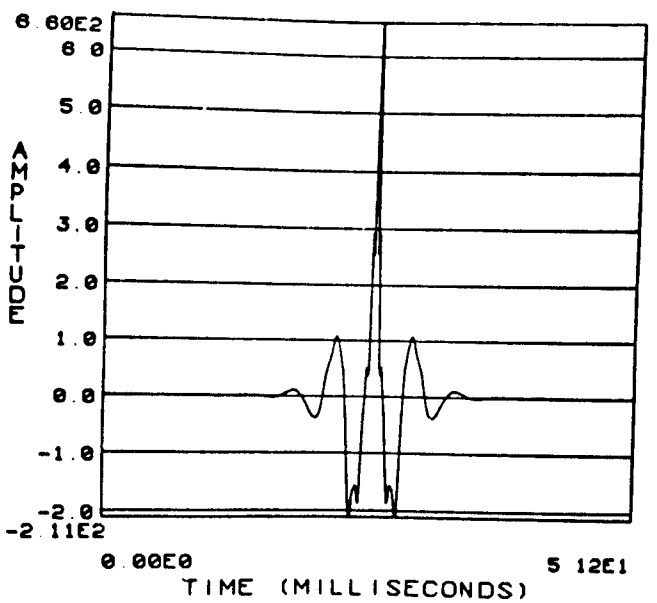

Figure 30. The zero phase impulse response.

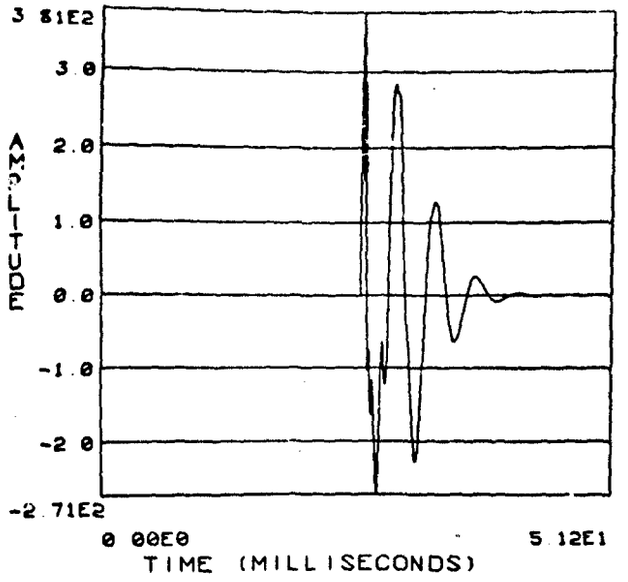

Figure 31. The minimum phase impulse response.

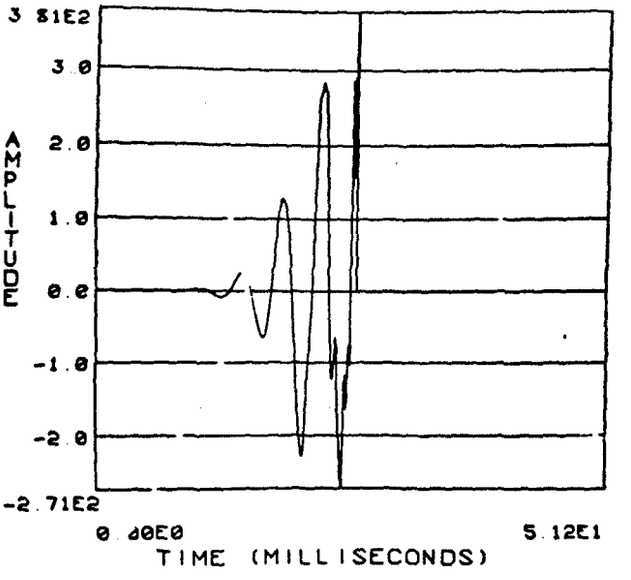

Figure 32. The maximum phase impulse response.

4.2 Frequency Extrapolation

The mechanical recording mechanisms used to produce acoustic disc recordings were incapable of reproducing signals with a frequency greater than about 4 K Hz. There is, however, considerable high frequency noise energy above 4 K Hz on these early records which can be attenuated by lowpass filtering. Thus, when the noisy signal was lowpass filtered at 4 K Hz, no loss of singing voice signal was experienced. Since this lowpass filtered signal contains no information above 4 K Hz, it cannot produce natural quality synthetic recordings.

An approximation of the missing high frequency information was derived by examining the high frequency responses of a singing signal. Figure 33 shows the log magnitude spectrum of a modern digital recording filtered at 15 K Hz and sampled at 35.0 K Hz. This spectrum has approximately a 6 dB per octave roll of from 4 to 10 K Hz. An attempt is made in this synthesis to provide each impulse response with a corresponding synthetic high frequency spectrum that rolls off a 6 dB per octave in the interval from 4 to 10 K Hz.

The synthetic high frequency spectrum is produced by symmetrically concatenating the log magnitude spectrum of the impulse response with zeros, thereby increasing its length four fold. After this alternation, the highest frequency represented in the log magnitude spectrum is

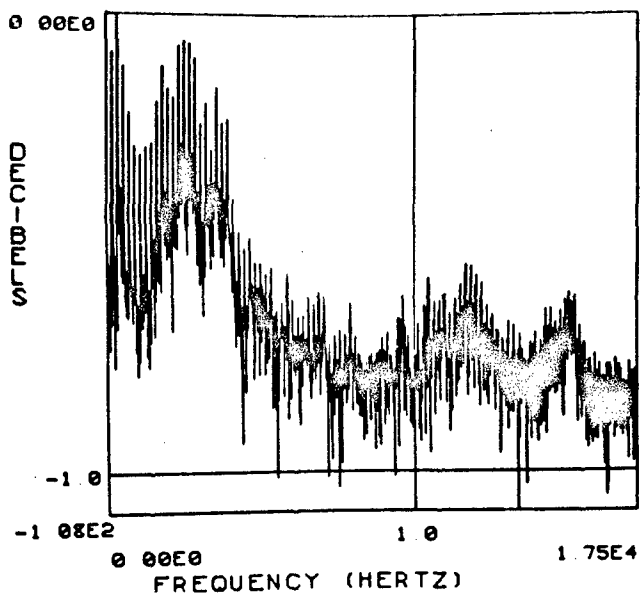

Figure 33. The log magnitude spectrum of a singer singing the $\overline{A}$ in $\overline{A}$braham. (This estimate was made utilizing 8192 samples of the signal lowpass filtered at 15 K Hz and sampled at 35 K Hz.)

20 K Hz. The frequency coefficients above 4K Hz are replaced with values on a line of slope 6 dB octave that passes through a point just above the 4 K Hz frequency.

Symmetry considerations imply that the phase of the log magnitude spectrum is zero at 5 K Hz. The phase of the synthetic high frequency impulse response is obtained by concatenating with zeros all phase coefficients above 5 K Hz.

The log magnitude spectrum is exponentiated, the real and imaginary parts of the complex spectrum computed from the magnitude and phase, and an inverse discrete Fourier transform applied. The result is an impulse response which is represented as a signal function sampled at 40 K Hz. Figure 34 shows the log magnitude spectrum of a section of the synthetic signal resulting from this process.

4.3 Interrelating Impulse Responses

Accurate estimation of impulse responses in the presence of non-stationary noise with unknown statistics at first seems an insoluble problem. However, since the vocal tract changes positions relatively slowly, its impulse response over each sampled interval is related to the impulse responses over adjoining intervals. The vocal tract was modelled as a time invariant linear system over each interval; it is, therefore, reasonable to expect that the log magnitude spectrum of the impulse response over each interval should be closely related to the log magnitude spectrum of the impulse responses over adjoining intervals.

A graphical representation of this relationship is obtained by plotting a surface of log magnitude spectra with time along the x-axis,

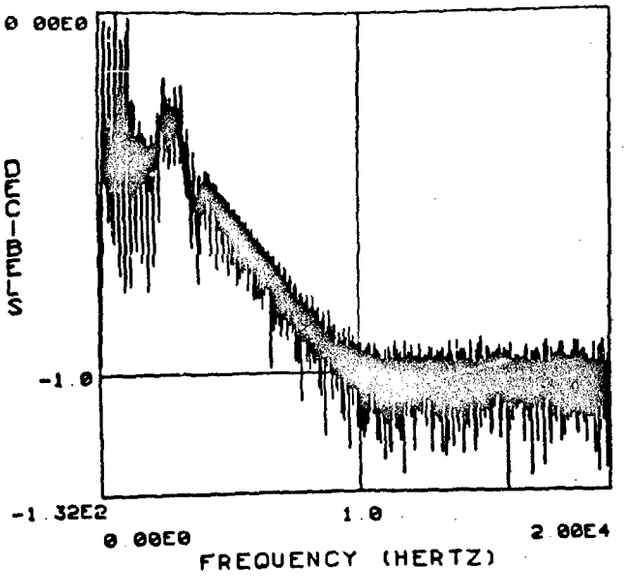

Figure 34. The log magnitude spectrum of a section of synthetic signal with artificial high frequency response. This is a in "Recitar!"

frequency along the y-axis, and log magnitudo spectra along the z-axis. Such a surface is plotted in Figure 35 for a section of noise free signal. A second such surface is plotted in Figure 36 for a section of singing voice signal in the presence of musical accompaniment and surface scratch. The surface generated from the noisefree signal is more regular than that from the noisy signal indicating that musical accompaniment and surface scratch cause significant variations of an impulse response's log magnitude spectra from those of its neighbors.

Three techniques were implimented to constrain the sequence of impulse responses and maintain their interrelationships. The first method smooths the graphically defined log magnitude spectral surface along the time axis for each frequency coefficient. The second technique utilizes a long time domain data window. A third procedure also found to maintain the impulse response interrelationships was to widen the shortpass lifter to include the cepstral peak used for pitch detection.

4.3.1 Log Spectral Smoothing

To smooth the log magnitude spectrum of each frequency coefficient, the step size for which a new impulse response of the vocal tract is estimated was selected to be 6.4 milliseconds and the data window width was set at 25.6 milliseconds. The cepstral liftering process was performed for each step up to the computation of the log magnitude spectrum of the impulse response. The one sided log magnitude spectrum and phase were stored until all the log magnitude spectra and phases were computed for the selection. The log magnitude values for each of

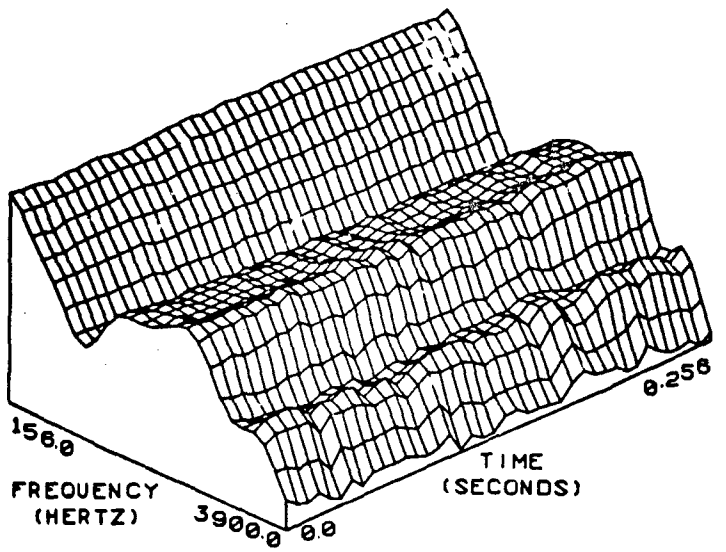

Figure 35. A surface of log magnitude spectra for a noisefree signal. This is the Ā sound in "Ābraham."

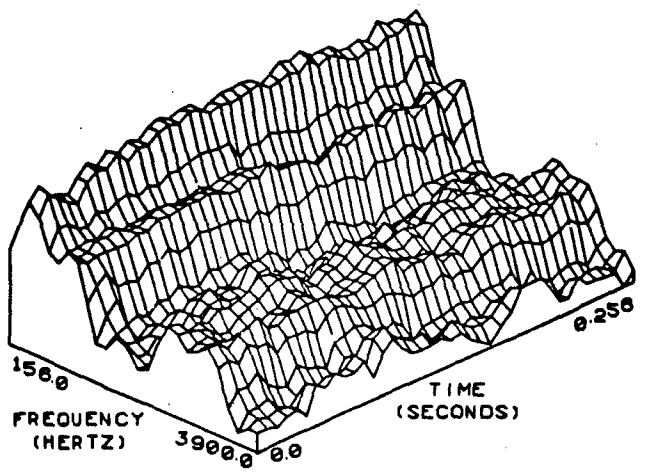

Figure 36. A surface of log magnitude spectra for a singing voice signal is the presence of musical accompaniment and surface scratch. This is the ā sound in "Recitar!".

the frequency coefficients along the time line were then convolved with a 16 point, zero phase, impulse response of a lowpass filter. These filtered log magnitude spectral values were then used with their corresponding phases to produce the sequence of impulse responses. Figure 37 shows the result of this process applied to the log magnitude spectra shown in Figure 36.

4.3.2 Increased Time Window Widths

The second system of constraint utilized long data windows. Lengthening the data segment used to estimate the vocal tract impulse response over each interval increases the amount of data which is shared with segments used for estimations in adjacent intervals. Impulse response estimates obtained from data segments with shared data will maintain their interrelationships despite the noise and surface scratch. Thus, by estimating responses for each 6.4 millisecond interval and using a 25.6 millisecond window, 75% of the data incorporated in a particular estimation will also be used in the estimations over its neighboring 6.4 millisecond intervals. This overlap is illustrated in Figure 38. Increasing the window width to 102.4 milliseconds, increases the overlap to 93.75% as illustrated by Figure 39. The beneficial effect of this overlap is somewhat reduced by the application of the Hanning window which weights the estimation segment heavily towards its center samples, but the interrelationships of the impulse responses from adjacent intervals is still present.

In addition to constraining consecutive impulse responses to be similar to their neighbors, the use of long data windows has two additional

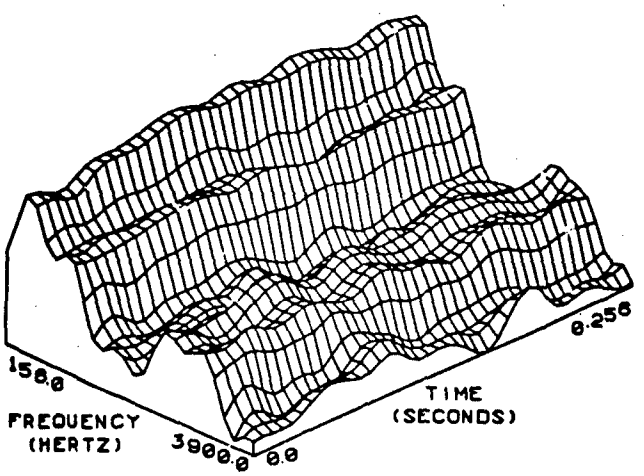

Figure 37. The smoothed log magnitude spectra of Figure 36. The log magnitude spectra of Figure 36 have been lowpass filtered along the time line for each frequency coefficient.

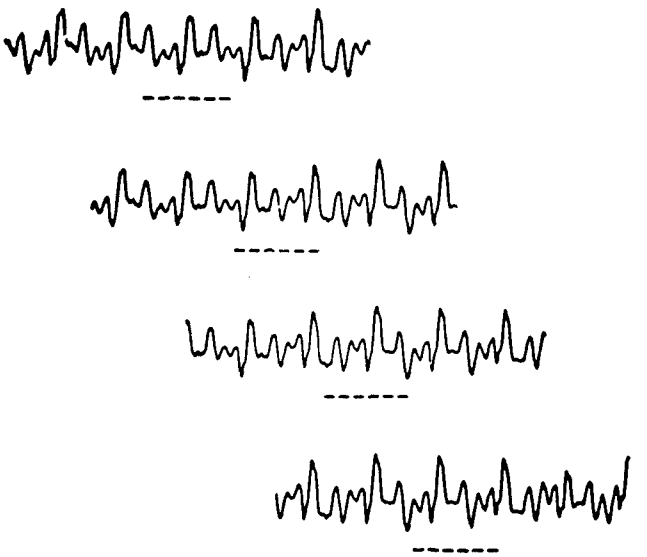

Figure 38. Four sections of data used for estimation of consecutive impulse responses. Each data segment is 25.6 milliseconds long for the estimation of the impulse response over the underlined 6.4 millisecond interval.

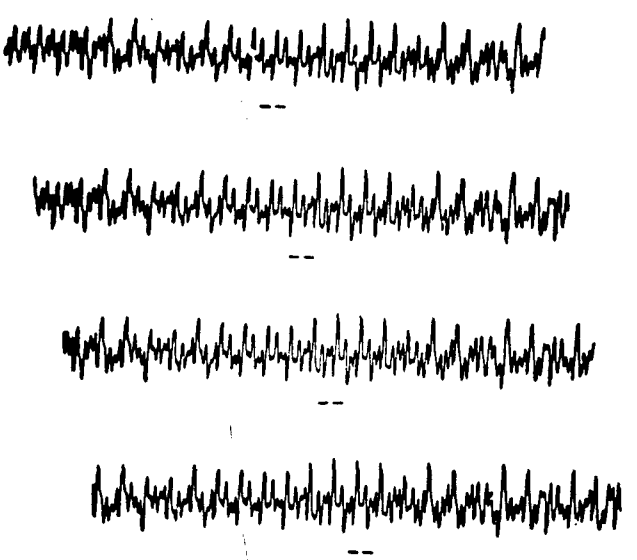

Figure 39. Four sections of data used for estimation of consecutive impulse responses. Each data segment is 102.4 milliseconds long for the estimation of the impulse response over the underlined 6.4 millisecond interval.

positive effects. First, the longer windows allow greater resolution of the log magnitude spectrum obtained from them. This increased resolution provides more information than an interpolation of an estimate made from less samples. Secondly, at each step of the estimation procedure, more samples are processed, consequently, the resulting impulse response estimation is longer than one obtained with a shorter window. Increasing the window length permits the vocal tract a longer memory, thereby partially overcoming the restriction of a finite impulse response.

There are major draw backs to the use of long data windows. First, the assumption of stationarity is sometimes violated, and second, the cost of computation is increased. Even though a new impulse response is estimated over each 6.4 millisecond interval, a long data window could blur events that happen in a short time period by spreading their energy over the entire interval. For instance, the energy of stops and plosives that occur in speech would be lost over a wide interval. The computational cost of computing the discrete Fourier transform used in this part of the process dwarfs the costs of any other computations in this analysis. Therefore, the cost of computing this complex part of the procedure determines the cost of the overall process. The present cost of computing a DFT using the Fast Fourier Transform algorithm [17] is on the order of

$$Cost_N = N*\log(N) * CF \text{ seconds}$$

where N is the number of samples in the transform and the computation factor, CF, is a machine dependent constant involving the time required

to perform a sequence of complex additions and multiplications. For an N of 256, the cost is

$$Cost_{256} = 256*\log_2(256) * CF = 256*8*CF = 2048 * CF$$

while for an N of 1024,

$$Cost_{1024} = 1024*\log_2(1024) * CF = 1024*10*CF = 10240*CF$$

Thus the cost of computing impulse responses for 102.4 millisecond windows is about five times as great as the cost of computing for 25.6 millisecond windows.

Hammett in his work on an adaptive spectrum analysis vocoder, has examined the selection of data window widths in detail [18]. The solution Hammett used to optimize the trade off between stationarity and stability of estimates was to utilize different length data windows for different intervals, depending upon a measure of the rate of change of parameters describing the signal. Thus, his vocoder uses a long window over selections of voice signal where the vocal tract changes little and short windows when it changes rapidly. Although this technique has proven extremely useful in speech, it has not been implemented in our filtering scheme since singing voice signals are composed mostly of long sections where the vocal tract changes are small.

4.3.3 Pitch Sychronous Reverberation

The third constraint implemented is one which is unique to this research effort. It is called pitch synchronous reverberation and is fundamentally different from what is classically accepted as reverberation. Whereas reverberation usually refers to the effect of resonances of

chambers or rooms, which are generally one or two second effects that account for an echoing quality of sound, pitch sychronous reverberation is a phenomena created artifically and lasts only a few milliseconds. The purpose of pitch synchronous reverberation is to constrain sequential impulse resonses rather than to simulate room reverberations. It is achieved using wide cepstral lifters in the impulse response estimation process.

Widening the cepstral lifter to include the cepstral pitch peak has not been used in speech research for two reasons. First, the separation of excitation and articulation functions by the convolutional model of speech production described in Chapter II relies on the notion that the cepstral pitch peak is due primarily to the excitation function rather than to the impulse response. The second and most important reason is that the previous bandwidth reduction studies were primarily concerned with finding the smallest set of parameters necessary to regenerate transmitted speech intelligibly. In those studies, the low quefrency cepstral values, or equivalently, some description of the smoothed log magnitude spectrum, were found to adequately characterize the vocal tract impulse response. Inclusion of additional parameters was found to increase the bandwidth.

A cepstrum derived in the classical manner described earlier, contains additionally combined components which represent the excitation function and the impulse response of the vocal tract. That is,

$$c(t) = v(t) + e(t)$$

It has been mathematically postulated and empirically demonstrated that low quefrency components contain much of the information describing the impulse response of the vocal tract [9]. Information about the impulse response may also be contained in the higher quefrencies. Thus, increasing the width of the short pass lifter in the cepstrum ought to increase the information about the impulse response of the vocal tract. If this increased width were to cause the inclusion of the cepstral pitch peak, one would expect considerable distortion in the impulse responses estimated. We have found this to be the case. However, the distortion caused by the inclusion of this peak manifests itself as a phenomenon we call pitch synchronous reverberation. This pitch synchronous reverberation has been found to be an effective tool for constraining the impulse responses to be similar over neighboring intervals.

The conventional short pass liftering process used to estimate impulse responses obtains estimates which have high amplitude only over a short time interval. This was previously shown in Figure 30. When several different impulse response estimates are jointly used to synthesize a segment of artifical signal over a constant pitch period interval, the result is:

$$s(t) \sum_{i=1}^{n} R_i(t - i \cdot \tau)$$

In this equation n is the number of impulse responses under consideration, s(t) is the resulting synthetic signal, $R_i(t)$ is the ith impulse response, and τ is the pitch period.

This convolution is illustrated in Figure 40 for three consecutively estimated impulse responses. Over short intervals, typically less than the pitch period, the synthetic signal is derived primarily from only one of the impulse responses. This occurs because the estimated impulse responses only have high amplitude over a short interval. If successive impulse responses differ significantly from one another in a random fashion, the resulting total signal will have random characteristics.

Suppose that before computation of the total signal each impulse response were convolved with a train of symmetrically decaying impulses spaced at the pitch rate as shown in Figure 41. The result of this convolution would be a modified impulse response. An example of the modified impulse response obtained by convolving the zero phase impulse response shown in Figure 30 with the train of impulses shown in Figure 41 is shown in Figure 42. This modification process has been defined as pitch synchronous reverberation. Its effect is to generate reduced copies of the high amplitude portion of the impulse response spaced at multiples of the pitch period. An example using impulse responses modified by this pitch synchronous reverberation to generate a synthetic signal for three impulse responses is shown in Figure 43.

The resulting time domain synthetic signal displays little evidence of the modification or distortion, since the modification was pitch synchronous. In addition, pitch syncronous reverberation averages the impulse responses used to derive the synthetic signal over the entire pitch interval. This reduces the random error which is characteristic of synthetic signals derived from only one or two of the highest

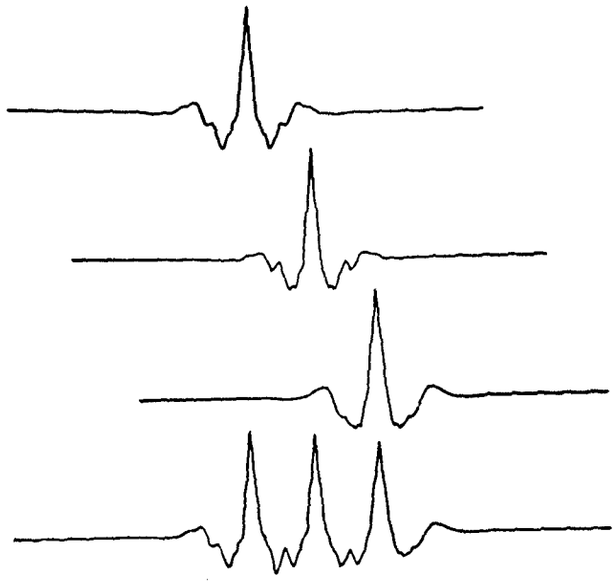

Figure 40. The convolutional process for three impulse responses obtained by conventional short pass liftering.

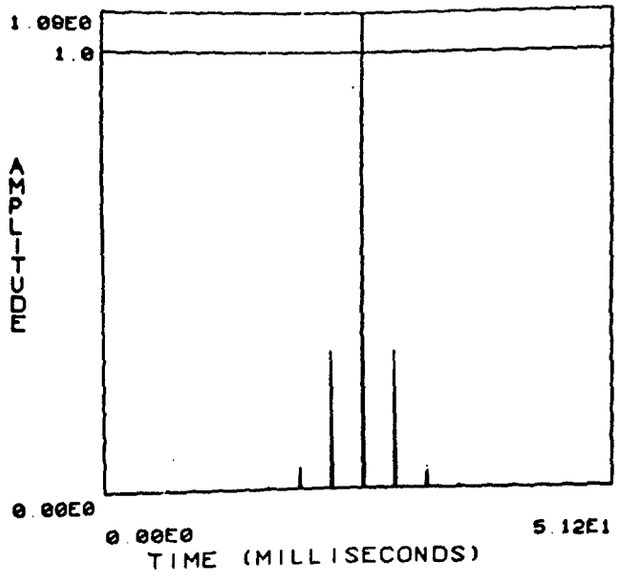

Figure 41. An example of a train of symmetrically decaying impulses spaced at the pitch rate.

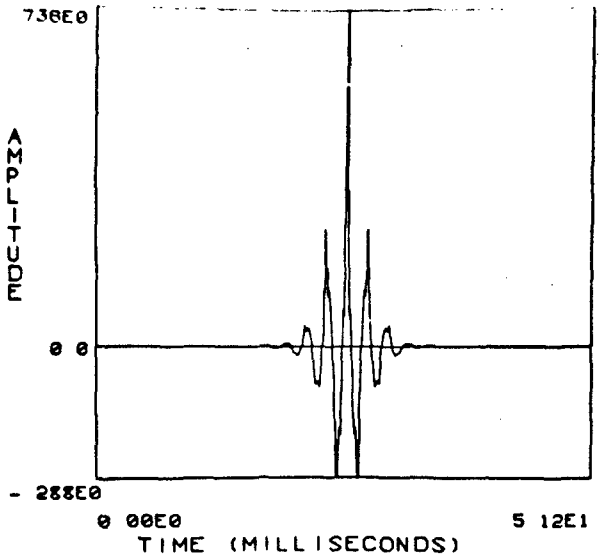

Figure 42. The convolution of the zero phase impulse response estimate shown in Figure 30 with the train of impulses shown in Figure 41.

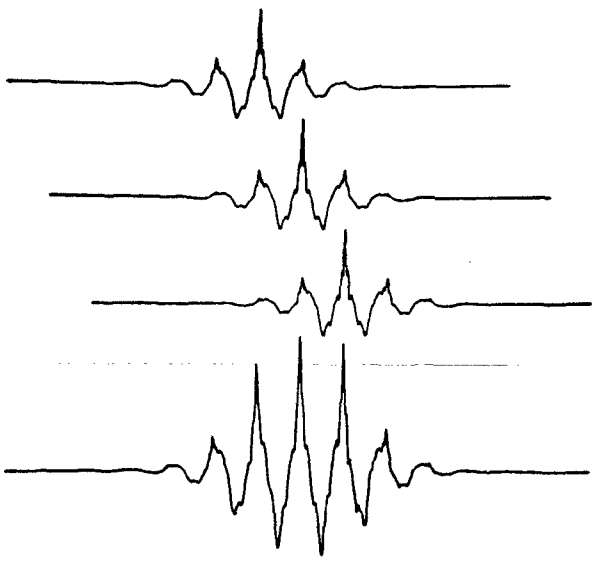

Figure 43. The convolutional process for three impulse responses which have been modified by pitch synchronous reverberation directly.

amplitude impluse responses.

Pitch synchronous reverberation can be effected directly for each impulse response, or it can be approximated for all impulse responses by including the cepstral pitch peak within their short pass lifters. To demonstrate this, we will show that the contribution of the cepstral peak to the impulse responses estimate is, indeed, a form of pitch synchronous reverberation.

An additive modification of the cepstrum manifests itself as a convolutional modification of the impulse response. Including the cepstral pitch peak alters the estimated impulse response by convolving it with a modifying impulse response derived from the cepstral peak. This modifying impulse response can be demonstrated to be a form of pitch synchronous reverberation. It will be referred to as the PSR (pitch synchronous reverberant) impulse response. The PSR impulse response can be determined by abstracting the pitch peak in the cepstrum as a pair of samples of amplitude H at the quefrencies with indices $\pm$ P. An example is shown in Figure 44.
The cepstrum is discrete and can be considered as a periodic sequence whose indices $\pm$ P are actually + P and (N - P).
Analytically,

$$cp(J) = H \text{ for } J = \pm P,$$
$$= 0 \text{ for } J \neq \pm P.$$

A cepstrum so defined for H = .3 and P = 32 is shown in Figure 44.
The log spectrum of the PSR impulse response is determined by computing

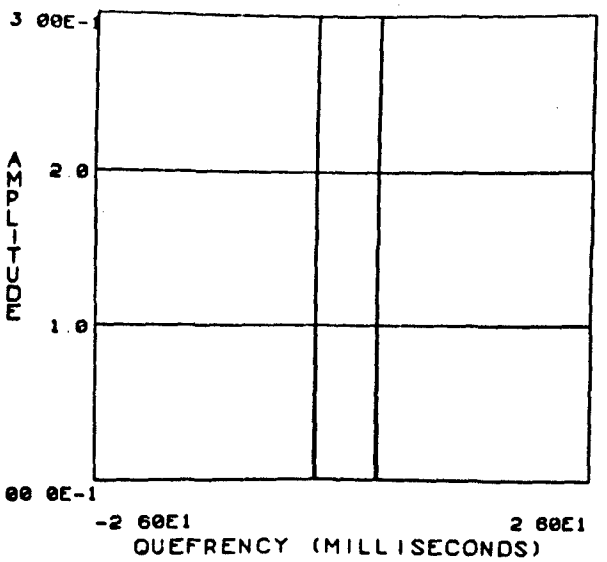

Figure 44. A Cepstrum with only non-zero values at the quefrencies 32. This is an approximation to the cepstral pitch peaks for a pitch of 313 Hz.

the discrete Fourier transform of this cepstrum.

$$\log sp_{PSR}(K) = DFT(cp(J)) = \sum_{J=0}^{N-1} cp(J).W_N^{-JK}$$

where $W_N = e^{j\frac{2\pi}{N}} = \cos(2\pi/N) + j \sin(2\pi/N)$.

therefore,

$$\log sp_{PSR}(K) = H.W_N^{-KP} + H.W_N^{KP}$$

$$= 2H \frac{e^{j(2\pi/N)KP} \quad e^{-j(2\pi/N)KP}}{2}$$

$$= 2H\cos((2\pi/N)KP)$$

This function, for H = .3 and P = 32 is shown in Figure 46.

The PSR impulse response is then obtained by computing the inverse DFT of the spectrum PSR. Thus,

$$PSR(J) = (1/N) \sum_{K=0}^{N-1} sp_{PSR}(K) \; W_N^{JK}$$

$$= (1/N) \sum_{K=0}^{N-1} sp_{PSR}(K).\cos((2\pi/N)JK)$$

since $sp_{PSR}(K)$ is an even function. Proceeding,

$$PSR(J) = (1/N) \sum_{K=0}^{N-1} 2H \cos(\frac{2\pi}{N} KP) \quad \cos(\frac{2\pi}{N} JK)$$

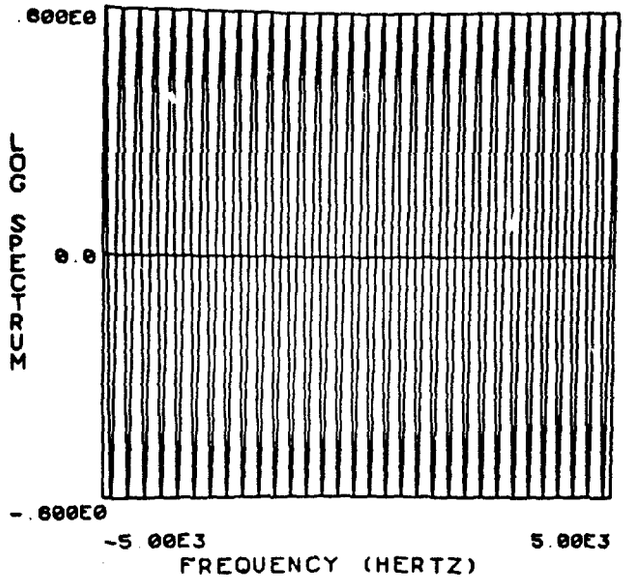

Figure 45. The logarithm of the spectrum associated with the PSR impulse response.

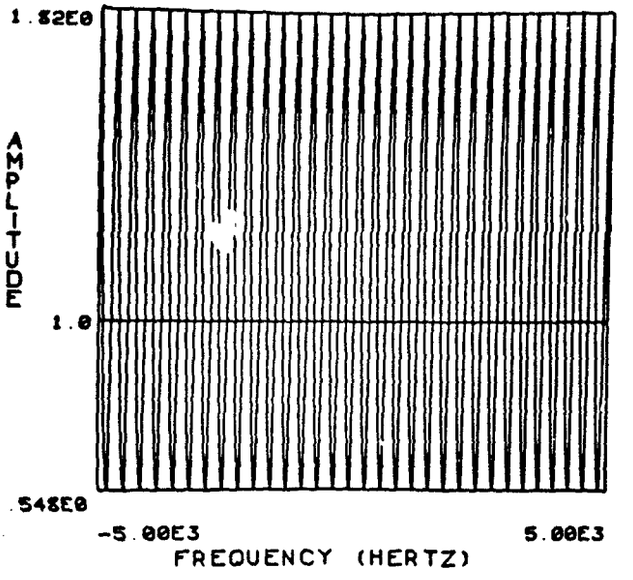

Figure 46. The spectrum associated with the PSR impulse response.

Because $e^{2H\cos((2\pi/N)t)}$ is even and periodic on the interval (0,N), it has a Fourier series expansion of the form

$$f(t) = \sum_{L=-0}^{\infty} F(L)\cos((2\pi/N)Lt)$$

where $$F(0) = \frac{1}{N}\int_0^N 2H\cos(\frac{2\pi}{N}t)\,dt$$

and $$F(L) = \frac{2}{N}\int_0^N 2H\cos(\frac{2\pi}{N}t)\cos(\frac{2\pi}{N}Lt)\,dt$$

This sequence of Fourier coefficients, though not analytically band-limited, has been found to decay rapidly for H less than one. The first 12 coefficients for an H of 0.5 were numerically computed using trapazoidal integration with a step size of N/4000. These are plotted in Figure 47. To show their decay more clearly, their logarithm was computed and plotted in Figure 48. For the purposes of this analysis, the coefficients with L greater than 10 have been assumed zero. Therefore,

$$PSR(J) = \frac{1}{N}\sum_{K=0}^{N-1}\sum_{L=0}^{10} F(L)\cos\frac{2\pi}{N}PLK)\ \cos(\frac{2\pi}{N}JK)$$

$$= \frac{1}{N}\sum_{L=0}^{10} F(L)\sum_{K=0}^{N-1}\cos(\frac{2\pi}{N}PLK)\cos(JA)$$

Now, since $\cos(A)\cos(B) = \frac{1}{2}\cos(A+B) + \frac{1}{2}\cos(A-B)$

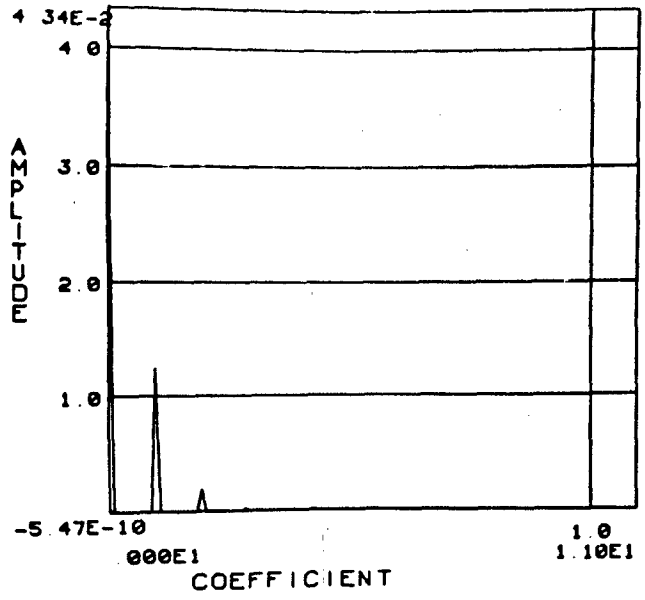

Figure 47. The first 12 coefficients in the Fourier series expansion of

$$e^{\cos u}$$

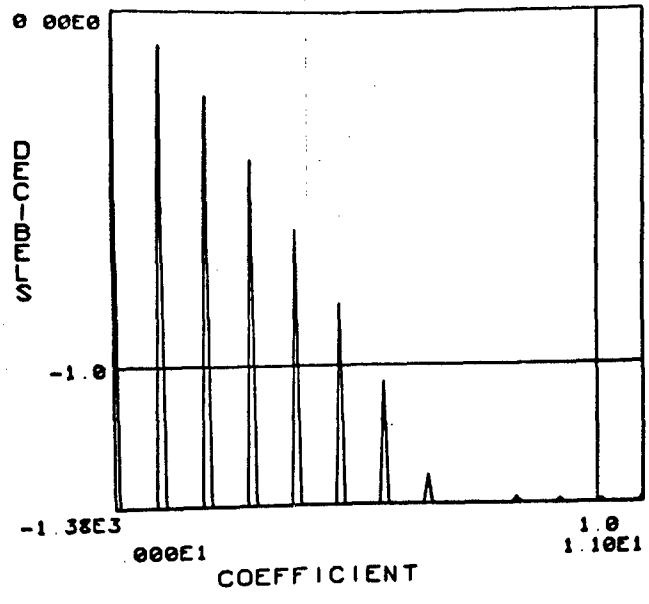

Figure 48. The logarithm of the first 12 coefficients in the Fourier series expansion of

$$e^{\cos u}$$

$$PRS(J) = \frac{1}{N} \sum_{L=0}^{10} F(L) \left[\frac{1}{2} \sum_{K=0}^{N-1} \cos\left(\frac{2\pi}{N}(PL-J)N\right) + \frac{1}{2} \sum_{K=0}^{N-1} \cos\left(\frac{2\pi}{N}PL+J)K\right) \right] \quad (4)$$

A proof is given in Appendix A which shows that if N is a power of 2, then

$$F(M) = \sum_{K=0}^{N-1} \cos \frac{2\pi}{N} MK = \emptyset$$

for all integers M that are not multiples of N or zero. Therefore, sums in Equation 4 will be nonzero only when LP-J or LP+J are multiples of N. When this is the case, each sum computes to N. Since $1 \leq P \leq N-1$, $0 \leq J \leq N-1$, and $0 \leq L \leq 10$, the condition that LP-J or LP+J are multiples of N occurs only for J multiples of P. Thus,

$$PSR(J) = F(L) \quad \text{for } J = LP$$

$$\text{negligible for } J \neq L$$

The negligible values for PSR(J) when J is not a multiple of P is necessary when the effects of higher order coefficients of F(L) are considered, but we have found that these values are not computationally significant. A PSR(J) function for H = .3 and P = 32 is shown in Figure 49.

The resulting PSR impulse response thus meets the requirements of being nonzero only at multiples of the pitch period and being decaying.

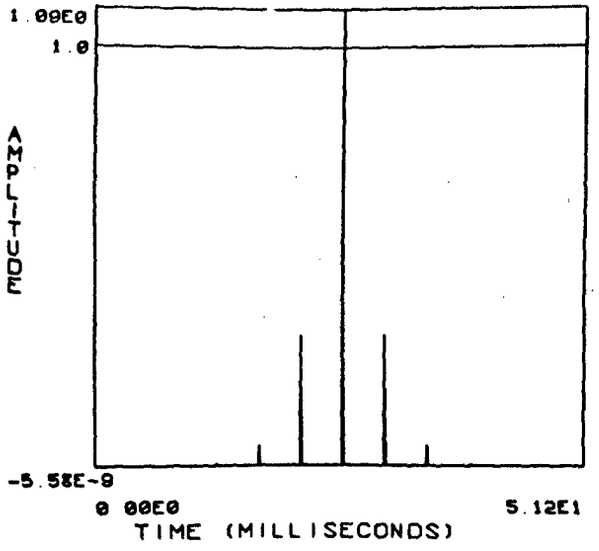

Figure 49. The effective PSR impulse response for H = .3 and P = 32.

The effect on the estimated impulse response of the inclusion of the cepstral pitch peak has been found to be an approximation to pitch synchronous reverberation.

4.4 Evaluation

The application of the homomorphic vocoder as a filter for singing signals requires a determination of the excitation function and an estimation of the impulse response before attempting to synthesize the singing signal without reproducing the overlaying noise. The impulse response was estimated by the use of classical short pass liftering in this analysis.

This filtering-synthesis application of the homomorphic vocoder had to overcome several unique problems when estimations of the impulse responses were made. First, the originally recorded singing signal had contained no frequency information in the 4 to 10 K Hz range. An attempt was made to approximate these missing high frequencies. The original reproduction was distorted not only by archaic recording equipment, but also with a variety of surface scratch noise and poor quality musical accompaniment. The true impulse response of the singing signal had to be estimated out of this noisy environment. Various constraints were employed during impulse response estimation to avoid incorporating the noise components.

Although short pass liftering is a well defined technique for estimation of impulse responses, there exists diversity in the phase accompanying the synthesized signal.

The subjective effect of different phases generated with synthetic speech has been extensively studied. Most researchers have concluded that minimum phase is most natural [19]. In the present filtering-synthesis analysis, degradations due to the presence of the background accompaniment and surface scratch mask most of the distinctions between different phased synthetic segments. The variations generated by using different phases were minor compared to the differences resulting from the modification of other parameters such as window width or the inclusion of pitch synchronous reverberation (PSR). This study, therefore, dwelt more heavily on the modifications caused by window width and pitch synchronous reverberation rather than those resulting from phase changes. The zero phase impulse response was utilized throughout this analysis and found computationally efficient as well as satisfactory for synthesis. Further analyses which deal with cleaner signals may wish to more extensively examine the subleties of phase changes.

Another modification of the impulse response estimation process was frequency extrapolation. The attempt to provide the articulation function with artificial high frequencies by linear extrapolation above 4 K Hz proved unsuccessful. The inadequacy of the linear extrapolation process was a result of two factors. First, there exists a natural cutoff in the vocal tract articulation function at 4 K Hz as illustrated by Figure 33, and second, there appears to be important singing signal information in the spectrum above 4 K Hz.

Linear extrapolation of the articulation function beginning at 4 K Hz (see Figure 34), results in an articulation function with

unnaturally high energy in the frequency range from 4 to 6 K Hz. Inclusion of these high energy frequencies manifests itself as an occasional crackling in the synthetic signal rather than supplying the missing signal information.

Except for crackling, no noticeable difference can be heard between the synthetic signal containing linear extrapolated artificial high frequencies and the signal filtered at 4 K Hz. In contrast, modern recordings compared before and after 4 K Hz filtering show noticeable degradation in the filtered signal. Therefore, it can be concluded that the important information above 4 K Hz is much more complex than that approximated with a simple linear extrapolation.

The missing high frequency information might be approximated in the following manner. A data base containing the high and low frequencies associated with phonemes could be compiled. The synthetic signal's low frequency articulation function would be compared for parallelism with the stored low frequency data. The matched stored low frequency data also has associated high frequency information. This high frequency information could then be used as the approximation for the synthetic signal's missing high frequencies.

It is also possible that there exists some constant relationship between the lower and high frequency portions of the articulation function that would allow a more accurate extrapolation. Investigations of both these alternate methods for providing high frequency information to the synthetic signal are massive projects and must be left to future research. However, a more successful method for compensating for the lack of high frequencies in the original signal need to be developed.

Several successful innovations were implemented during estimation of the impulse responses. Each impulse response estimate was originally made in the presence of different background accompaniment and surface scratch resulting in inaccurate impulse response estimates. The synthetic signal generated directly from these inaccurate estimates had random characteristics. This randomness was reduced by constraining and modifying the impulse response estimates to be related to one another.

The first and most direct constraint introduced was smoothing the impulse estimate's log magnitude spectra, or articulation functions, along the time axis.

Low pass filtering the time history of each frequency coefficient accomplished this smoothing. Although the synthetic signals derived from these constrained articulation functions were much less random, the clarity of the synthetic signal produced was slightly distorted when the articulation function was extremely constrained.

The time histories of the frequency coefficients were low pass filtered using seven different filters. These filters allowed the coefficients to have cutoff frequencies varying from 9.8 cycles per second to 78.3 cycles per second in steps of 9.8 cycles per second. The synthetic signal generated from the most constrained articulation functions, those with cutoffs of 9.8 cycles per second, was distorted because its articulation function sequence failed to change rapidly enough. The synthetic signal generated from the least constrained, nonfiltered articulation functions, those with cutoffs of 78.3 cycles per second,were dominated by the random characteristics of the

erroneous impulse response estimates. Utilizing a filter with a cutoff at 29.4 cycles per second, intermediate between the extremes of most constrained and nonconstrained, produced the highest quality synthetic signal.

A second and more noticeable constraint was the introduction of pitch synchronous reverberation (PSR) which resulted from including the cepstral pitch peak in the impulse response estimation. Not only was more cepstral information about the impulse response retained for the estimate, but the effect of cepstral peak inclusion is to distribute the energy from each estimate over a wider time interval. A cepstral window with a width of six times the quefrequency just below the cepstral pitch peak yielded a most acceptable final synthetic signal.

There is one drawback to pitch synchronous reverberation. A pitch sychronous noise perceived as a high pitched background sound occasionally is introduced into the synthetic signal. This effect occurs if the secondary impulse images do not align with neighboring impulse responses when the pitch rapidly changes. This effect seldom occurs, is muted, and is less objectionable than the randomness of the impulse response estimates occuring in the absence of PSR.

The final constraint imposed on the impulse response estimates was achieved by selection of overlapping data windows for each estimation. Three different window widths were used for the impulse response estimates corresponding to three different percentages of overlap. The amount of overlap varied from a low of 75% using the 25.6 millisecond window to a high of 93.75% using the 102.4 millisecond windcw. The estimates were least constrained using the 25.6 millisecond

window. Utilizing this size data window, the synthetic signal had a random background signal associated with it. This introduced background signal was due to randomness between successive impulse response estimates. The estimates were most constrained using the 102.4 millisecond window. Utilizing this size data window, the synthetic signal had hardly a trace of the random, background noise heard in the signal derived using the 25.6 millisecond window. However, another degradation appeared.

The synthetic signal sounded as if it had been recorded through a long tube. A short time echo had appeared. One explanation for this effect is that increasing the data window width has also increased the width of the estimated impulse response. This longer impulse response, coupled with the impulse images derived from the pitch synchronous reverberation effect, produced repeations of the impulse response which were separated in time and perceived as an echo. A 51.2 millisecond window was utilized as an ideal compromise. This window width was not long enough to allow a perceptable echo. However, it provided enough overlap (87.5%) to significantly reduce the random noise effect observed when synthesizing from data sampled with a 25.6 millisecond window.

Increasing the data window width substantially improved the efficiency of using pitch sychronous reverberation to avoid incorporation of noise components. This increased window width substituted for log magnitude smoothing as an effective constraint. Although the frequencies above 4 K Hz could not be successfully approximated by linear extrapolation, it was possible to synthesize the singing signal without

this information. Interrelating the impulse responses significantly aided the synthesis of the singing signal from its noisy environment.

V. SYNTHESIS

To obtain the synthetic voice signal, the estimated excitation function and the estimated sequence of impulse responses must be convolutionally combined. Several problems arose while implementing this combination peculiar to the digital nature of the homomorphic vocoder filtering process. Distortion was introduced into the synthetic signal when the impulse response estimates were derived from data which had been improperly windowed. The choice and utilization of the data window is, therefore, extremely important. High resolution pitch information is provided by the homomorphic vocoder. However, it was necessary to quantize some of this pitch information in order to reduce the computational expense during processing.

Another problem encountered in this application of the homomorphic vocoder involves identification of intervals which, due to the high energy of the noise signal, were not marked as voiced signals. A large number of these intervals occur at the beginnings and endings of voiced segments. Synthetic signals which included estimates from these sections were characterized by inclusion of orchestral signals and abrupt starts and stops. The homomorphic vocoder produces a synthetic signal which is directly transferred onto analog tape.

Any singing signal so produced lacks the room acoustical modifications incorporated in signals produced in recording studios. These room acoustical modifications soften the singing signal and are an

integral part of the total recorded sound identified as a modern reproduction. To enhance the synthetic signal and more accurately produce a reproduction as it might be recorded today, a procedure was implemented to provide the synthetic signal with artificial room acoustics.

5.1 Windowed Impulse Responses

The procedure for computing the estimates of the impulse responses is essentially correct, however a slight problem exists when these estimates are directly used to generate a synthetic voice signal. Because of the discrete nature of the impulse response estimation, the estimated impulse response is periodic. However, the actual discrete result of the inverse Discrete Fourier transform of the articulation function is equally spaced samples only over one period of this periodic function. If this information is used directly in the synthesis process, it is equivalent to using the correctly estimated impulse responses multiplied by a Fourier window. This results in a distortion of the synthetic signal due to leakage by the Fourier window.

We have found Hanning windowing of each impulse response produces estimates with tolerable distortion.

5.2 Effect of Pitch Quantization

The resolution of the estimated lengths of the pitch period was 0.025 milliseconds during the determination of the excitation function.

This superresolution, four times greater than the sampling rate of the noisy signal, aids the excitation function separation program in detecting the slight variations characteristic of musical accompaniment. This high resolution pitch information can be directly utilized in the 10000 samples per second convolution process. However, this convolution is computationally expensive because it requires that the excitation function contain impulses which do not fall at integer multiples of the sampling rate.

For pitch rates where the period is an integer multiple of the sampling rate, the excitation function

$$e(t) = (\sin((2\ pi/10000)*(t-t_p))/((2pi/10000)*(t-t_p))$$

is only nonzero at one sample. In contrast, for pitch rates which are non integer multiples, this function is non negligible over many samples. It is the convolution of two sequences which are non-zero over several samples, which is computationally expensive. Although high speed convolution techniques have been developed [20], their computational expense is still considerable when there is more than one non zero sample in both sequences.

Computational efficiency was increased by quantizing the pitch intervals so that each impulse occurred at an integer multiple of the sampling rate. With the exception of some long, gradual glides, this pitch quantization did not noticeably effect the overall quality of the synthetic voice signal. The high resolution pitch information in the glide sections was recovered and utilized by a four fold

interpolation of their impulse responses. This interpolation yielded 40 K Hz sampled impulse responses which were then resampled with delays chosen to allow the effective period to contain the high resolution pitch periods. Computation of the 40 K Hz signals with 40 K Hz sampled impulse responses required no quantization of the pitch periods.

5.3 Corrections for Discontinuous Primary and Terminal Sections

The processes used for pitch detection and impulse response estimation in this research detect the pitch and articulation functions of the dominant signal. However, the estimated excitation and articulation functions from sections where the background accompaniment signal is greater than the singing signal has orchestral characteristics rather than those of singing voice. Estimates with these orchestral characteristics were easily detectable when the synthetic signal was auditioned. These fallacious estimates had to be deleted.

A simple method utilized to avoid incorporation of estimates made on orchestral accompaniment detects these estimates by their signal to noise ratio. Although the singing signal is 10 to 20 dB larger than the noise signal for most of the recording, those sections at the beginnings and endings of singing segments fall well below this high signal to noise ratio. Any voiced signal which has a high signal to noise ratio also has a cepstral peak detectable as coming from the singing signal. Estimates with accompaniment characteristics can be eliminated by only synthesizing the voice signal over intervals where its energy is high enough to produce detectable cepstral peaks.

This procedure, although effective in eliminating explicit traces of the accompaniment, causes the synthetic signal to have unnaturally discontinuous beginnings and endings. The required signal to noise ratio is reached only after the singing is well underway, and is lost well before the singing has ceased. Consequently, some primary and terminal sections of singing are not synthesized because their cepstral peaks were not detected. Both these omissions and the high noise energy at the beginnings and endings of singing segments causes the synthetic signal to have a jerky quality. The synthetic signal starts and stops abruptly.

Correct synthesis of beginning and ending segments requires excitation functions and impulse response estimates which are not adequately provided by the above procedures. Reasonable approximations of these functions can be determined from the excitation functions and impulse responses estimated over neighboring intervals, because the singing signal model is stationary over long intervals. To correct the problems of high energy orchestral accompaniment and the ensueing discontinuity of the primary and terminal estimates, we have 1) only computed singing voice signal estimates for the excitation functions and impulse responses over these intervals with a singing signal of higher energy than the orchestral accompaniment as implied by the presence of the correct cepstral pitch peak, and 2) utilized estimates from sections nearest beginnings and endings as approximations for the required functions over these primary and terminal sections. In addition, the approximated primary and terminal impulse responses are assumed constant over these sections and the excitation functions

are made to exponentially rise or decay. This provides the singing sections with a synthesized signal which is reasonably continuous from beginning to end and characterized by natural increases and decreases rather than abrupt starts and stops.

5.4 Artificial Room Reverberation

The synthetic signal produced in this analysis lacks reverberation and therefore sounds unlike a modern recording made in a reverberating studio. The synthetic signal is perceived as if it had been recorded in an anechoic chamber. The effect of room reverberation was artificially introduced into the synthetic signal by convolving it with an exponentially decaying set of impulses that have flat spectral characteristics.

Schroeder et al. have correlated some of an enclosure's subjectively desirable effects on sound with the physical characteristics of its reverberations [21]. These investigators have noted that rooms in which reverberations are equally attenuated in frequency and decay exponentially are desirable for sound reproduction. They also observed that the time required for the energy from a sound burst to decay to one one-thousandth of its initial value (-60 db) was optimally 1.8 seconds. A 1.8 second delay time or reverberation time was chosen for the artificially induced reverberations.

To introduce artificial reverberations with the desirable characteristics, the synthetic signal was convolved with a 1.6 second sample sequence of zeros and values from one to .001. This sequence contained twenty nonzero values occuring randomly throughout the 1.6

second interval. The nonzero values were chosen with exponentially decaying magnitudes whose decay constant was 1.8 seconds.

By assigning the sequence of impulses random polarity, its spectral characteristics became nearly flat. The number of nonzero values was restricted to twenty for the sake of computational efficiency since the fast convolution algorithm would become unmanageable for a 1.6 second interval. The 1.6 second interval could contain as many as 64,000 nonzero samples at the 40 K Hz sampling rate. After convolution with the reverberation sample sequence the synthetic signal is finally recorded on analog tape.

5.5 Evaluation

The synthesis process consisted of three consecutive stages. First, the derived excitation function and impulse responses were modified to incorporate natural beginnings and endings. Next, the excitation function was convolutionally combined with the appropriate impulse responses to form a reverberation-free synthetic signal. Finally, the synthetic signal was convolved with a 1.6 second impulse response chosen to simulate room reverberations.

Modification of the excitation function and impulse responses to effect natural rather than abrupt beginnings and endings was essential. These functions had been estimated from a noisy environment and failed to completely describe the actual singing signal information. The singing signal's amplitude did not exceed the noise and was not detected until about 50 milliseconds of the signal had commenced. The singing signal also dropped below detectable amplitude before the

singing had terminated. These missing intervals resulted in the abrupt beginnings and endings in the synthetic signal. To compensate for this deficiency, the first 50 milliseconds of silence immediately before and after each singing section was substituted with a synthetic singing signal. An excitation function for computation of these approximated intervals was supplied by using the excitation function from the closest adjacent detectable interval encountered in processing of the original singing signal. The impulse responses for these missing intervals were provided by using the impulse response from the closest detectable adjacent interval multiplied by a weighting function. This weighting function was exponential and decayed from a value of 1 to a value of $\frac{1}{1000}$ as it proceeded from the detectable interval to 50 milliseconds away. Whenever two consecutive singing segments were separated by less than 100 milliseconds of silence, this same technique for modification was performed. The impulse response of the first singing segment was multiplied by a function which decayed from one to one one thousand as it proceded to the halfway point of the silent interval. The impulse response of the second singing section multiplied by the exponential function was supplied for the silence from the halfway point to the second detectable interval. This modification improved the synthetic signal's tendency to begin and terminate with unnatural abruptness.

The selection of 50 milliseconds as the constant size of the silent beginning and ending intervals obscured by noise was the major shortcoming of this correction technique. An audition of the modified synthetic signal illustrated that these obscured intervals must

actually be extremely variable in length. A massive interactive processing would have been required to correctly estimate the different lengths of the obscured silent sections. This processing was not feasible for the present analysis and would not be desireable in a generalized technique for filtering by synthesis. It is possible that if the total time length of each segment could be determined accurately, some of the missing intervals could be supplied by auditioning.

The attempt at providing the synthetic signal with artifical room reverberation has been only partially successful. Complete success was not achieved for atleast two reasons. First, the number of nonzero elements in the impulse response of the room was restricted to twenth. The process of introducing room reverberations outlined above was not only applied using 20 nonzero elements, but also using 40 and 120 nonzero elements. In addition, the reverberation time was varied from 1.6 seconds to 0.8 seconds and 0.2 seconds, using impulse response lengths of the reverberation times. The most natural sounding results were obtained using a reverberation time of 0.2 seconds, 120 nonzero samples, and an impulse response length of 200 milliseconds. This corresponds to the highest density of nonzero elements utilized, indicating the necessity of using a large number of nonzero elements.

The second reason why only limited results were obtained is that the impulse response of the room has not been constrained

to be related to an acoustially desireable room. These two problems might be overcome simultaneously be estimating the impulse response of a subjectively desireable room and using this estimate to effect artifical room reverberation.

VI. CONCLUSIONS

The goal of this project has been to examine the homomorphic vocoder as a filtering tool. This goal has been accomplished. In this analysis, the homomorphic vocoder has been developed and modified to effect a highly successful filtering scheme. This has been accomplished by careful construction and examination of the homomorphic vocoder and by the development and refinement of constraints. In particular, the inclusion of pitch synchronous reverberation has been extraordinarily useful for achieving a natural synthetic voice signal.

The linear extrapolation of the articulation function into higher frequencies yielded results that indicated a more complex scheme is required to effect natural high frequency information. Interpolating the excitation and articulation functions into the initial and terminal sections of voiced segments provided an improvment in the quality of the synthetic signal. The notion of overlapping the data from which the impulse response estimates are made has been found to be an effective tool for interrelating adjacent impulse responses. Spectral smoothing provided the most direct constraint on the articulation functions; however, its subjective effect is less favorable than either the pitch synchronous reverberation or the overlapping segment constraints.

We conclude, therefore, that utilizing the constraints developed

in this analysis, the homomorphic vocoder can be used successfully as a filtering device for singing voice signals. The success of this project indicates that efforts into related filtering problems, such as filtering speech signals, may also be fruitful using the techniques developed in this analysis.

BIBLIOGRAPHY

1. A.V. Oppenheim and R.W. Schafer, "Homomorphic Analysis of Speech," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 221-226, June 1968.

2. V. Zue, "Translation of divers' speech using digital frequency warping," Research Lab. of Electronics, M.I.T., Cambridge, Mass., Quart. Progr. Rept. 101, pp. 175-182, April 15, 1971.

3. T.G. Stockham, Jr., "The application of generalized linearity to automatic gain control," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 267-270, June 1968.

4. H. Dudley, "Remaking Speech," *J. Acoust. Soc. Am.*, vol. 11, pp. 169-177, October 1939.

5. M.M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 262-266, June 1968.

6. A.M. Noll, "Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection," *J. Acoust. Soc. Am.*, vol. 36, pp. 296-302, February 1964.

7. J.P. Olive, "Automatic formant tracking by a Newton-Raphson technique," *J. Acoust. Soc. Am.*, vol. 50, pp. 661-670, February 1971.

8. R.W. Schafer and L.R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Am.*, vol. 47, pp. 634-648, February 1970.

9. A.V. Oppenheim, R.W. Schafer, and T.G. Stockham, Jr., "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, pp. 1264-1291, August 1968.

10. E.A. Guillemin, *Theory of Linear Physical Systems*. New York: Wiley, 1963, ch. 16.

11. B. Gold and C.M. Rader, *Digital Processing of Signals*. New York: McGraw-Hill, 1969, pp. 137-140.

12. A.M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, pp. 293-309, February 1967.

13. A.V. Oppenheim, "Speech analysis-synthesis system based on homomorphic filtering," *J. Acoust. Soc. Am.*, vol. 45, pp. 458-465, February 1969.

14. B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637-655, February 1971.

15. T.G. Stockham, Jr., "A-D and D-A converters: their effect on digital audio fidelity," Preprint, 41st Convention, Audio Engineering Soc., New York, October 1971.

16. M.R. Schroeder and A.M. Noll, "Recent studies in speech research at Bell Telephone Laboratories," *Proc. 5th Internatl. Congr. Acoustics*, Liege, 1965.

17. J.W. Cooley and J.W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, pp. 297-301, April 1965.

18. J.C. Hammett, Jr., "An adaptive spectrum analysis vocoder," Ph.D. Thesis, Dept. of Elec. Engrg., Georgia Inst. Tech., Atlanta, Ga., 1971.

19. B. Gold, "Experiment with speechlike phase in a spectrally flattened pitch-excited channel vocoder," *J. Acoust. Soc. Am.*, vol. 36, pp. 1892-1894, October 1964.

20. T.G. Stockham, Jr., "High-speed convolution and correlation," *Spring Joint Computer Conf., AFIPS Proc.*, vol. 28, pp. 229-233, 1966.

21. M.R. Schroeder, B.S. Atal, G.M. Sessler, and J.E. West, "Acoustic measurements in Philharmonic Hall (New York)," *J. Acoust. Soc. Am.*, vol. 40, pp. 431-434, February 1970.

APPENDIX

APPENDIX A. If N is a power of two, then

$$F(M) = \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N} Mk\right) = 0$$

for all integers M that are not multiples of N or zero.

Proof: First, we need only consider integers M which are in the interval (1,N-1) for if M is not in this interval, then

$M = M' \pm jN$

where M' is an integer within the interval and j in an integer.

Therefore,

$$F(M) = \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N}(M' \pm jN)k\right)$$

$$= \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N}M'k \pm \frac{2pi}{N}jNk\right)$$

$$= \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N}M'k \pm 2pi.jk\right)$$

$$= \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N}M'k\right)$$

Hence, F(M) = F(M'), where M' is in the interval (1,N-1).

Since N is a power of 2, N=2n for some integer n which is also a power of two. Therefore,

$$F(M) = \sum_{k=0}^{N-1} \cos\left(\frac{2pi}{N}Mk\right) + \sum_{k=0}^{2n-1} \cos\left(\frac{2pi}{N}Mk\right)$$

$$= \sum_{k=0}^{n-1} \cos\left(\frac{pi}{n}Mk\right) + \sum_{k=n}^{2n-1} \cos\left(\frac{pi}{n}Mk\right)$$

$$= \sum_{k=0}^{n-1} \cos\left(\frac{pi}{n}Mk\right) + \cos\left(\frac{pi}{n}M(k+n)\right)$$

$$= \sum_{k=0}^{n-1} \cos\left(\frac{pi}{n}Mk\right) + \cos\left(\frac{pi}{n}Mk + Mpi\right)$$

Since $\cos(A + B) = \cos(A)\cos(B) - \sin(A)\sin(B)$

and $\sin(Mpi) = 0$ for all integers M,

$$F(M) = \sum_{k=0}^{n-1} \cos\left(\frac{pi}{n}Mk\right) + \cos(Mpi)\cos\left(\frac{pi}{n}Mk\right)$$

$$= (1 + \cos(Mpi)) \sum_{k=0}^{n-1} \cos\left(\frac{pi}{n}Mk\right)$$

If M is an odd interger, then, since $\cos(Mpi) = 0$ for all odd integers M, $F(M) = 0$, and the theorem holds. If M is an even integer, then $M = 2m$ for some integer m. Hence

for even M,

$$F(M) = 2. \sum_{k=0}^{n-1} \cos(\frac{pi}{n} 2mk) = 2. \sum_{k\ 0}^{n-1} \cos(\frac{2pi}{n} mk)$$

But this is just the original equation. Since n is also a power of 2, the above reasoning can be applied recursively. Since after each recursion, $1 \leq m \leq n-1$, m will become odd before n, and F(M) = 0, for all integers M not multiples of N.